National Park Service

National Park Service
U.S. Department of the Interior

Sonoran Desert Network
Data Management Plan

October 2005

Deborah L. Angell
Data Manager

National Park Service
Inventory and Monitoring Program
Sonoran Desert Network
7660 E. Broadway Blvd.
Tucson, Arizona 85710-3776

SONORAN DESERT NETWORK
Inventory and Monitoring Program

Recommended citation:

Angell, Deborah L. 2005. Sonoran Desert Network Data Management Plan. National Park Service, Inventory and Monitoring Program, Sonoran Desert Network, Tucson, AZ. 80 pp.

Initial distribution:

SODN website: http://www1.nature.nps.gov/im/units/sodn/data.htm

Revision history log:

Previous Version	Date of Revision	Author	Changes Made	Reason for Change	New Version Number

Table of Contents

List of Tables

List of Figures

Acknowledgements

Much of the material in this data management plan was developed through the collaboration of a hardworking group of data managers and others. I would like to acknowledge and extend my appreciation to Margaret Beer (Northern Colorado Plateau Network), John Boetsch (North Coast and Cascades Network), Rob Daley (Greater Yellowstone Network), Fred Dieffenbach (Northeast Temperate Network), Patrick Flaherty (Appalachian Highlands Network), Teresa Leibfreid (Cumberland Piedmont Network), Bill Moore (Mammoth Cave National Park), Dorothy Mortenson (Southwest Alaska Network), Velma Potash (Cape Cod National Seashore), Gareth Rowell (Heartland Network), Geoffrey Sanders (National Capital Region Network), Sara Stevens (Northeast Coastal and Barrier Network), Doug Wilder (Central Alaska Network), and Mike Williams (Heartland Network).

A special thank you to John Boetsch, who agreed to take on the tremendous responsibility of coordinating this group. John kept us on the right track throughout this process, and I doubt we would have accomplished as much as we did without his commitment.

Thanks to Lisa Nelson, Joe Gregson, and Steve Fancy for their comments and suggestions on initial drafts of the individual plan chapters.

Finally, I appreciate the contributions network staff and cooperators Andy Hubbard, Theresa Mau-Crimmins, Natasha Kline, Emily Dellinger, Jason Welborn, Bill Halvorson, and Brian Powell.

Executive Summary

> *Data and information are the basic products of scientific research. In ecological research, where field experiments and data collections can rarely be replicated under identical conditions, data represent a valuable and, often, irreplaceable resource . . . In long-term ecological studies, retention and documentation of high quality data are the foundation upon which the success of the overall project rests.*
>
> *Brunt 2000*

Information is the common currency among the many different activities and people involved in the stewardship of National Park Service (NPS) natural resources. As part of the Service's effort to "improve park management through greater reliance on scientific knowledge," a primary purpose of the Inventory and Monitoring (I&M) Program is to develop, organize, and make available natural resource data and to contribute to the Service's institutional knowledge. The I&M Program's efforts to identify, catalog, organize, structure, archive, and disseminate relevant natural resource information will largely determine the Program's efficacy and image among critics, peers, and advocates.

The NPS is a highly decentralized agency with complex data requirements. The primary audience for many of the products from the I&M Program is at the park level – providing park managers with the information they need to make better-informed decisions and to work more effectively with other agencies and individuals for the benefit of park resources. However, certain data are also needed at the regional or national level for a variety of purposes, and as stated by the National Park Advisory Board (2001), the findings "must be communicated to the public, for it is the broader public that will decide the fate of these resources."

The Sonoran Desert Network (SODN) Data Management Plan presents the overarching strategy for ensuring that I&M Program data are documented, secure, accessible, and useful for decades into the future. The plan refers to other guidance documents, standard operating procedures, and detailed protocols that convey more specific standards and steps for achieving our data management goals. The plan acts as a foundation upon which to build as new protocols are developed, advances in technology are adopted, and new concepts in data management philosophy are accepted.

Data and Data Management: An Overview

Collecting natural resource data is a critical step toward understanding the structure and function of the evolving ecosystems within our National Parks. We analyze and synthesize these 'raw' data to model various aspects of ecosystems. In turn, we use our results and interpretations to make management decisions about the Park's vital natural resources. Thus, *data* collected by researchers and managed by the Sonoran Desert Network according to this Plan is transformed into *information* through analyses, syntheses, and modeling.

Any good set of data – whether collected last week or 20 years ago – must be accompanied by enough explanatory documentation (*e.g.*, how and why it was collected) so that we can understand it and use it with confidence. Therefore, our Network data management system cannot simply attend to the tables, fields, and values that make up a data set. It must also provide a process for developing, preserving, and integrating the context that makes data interpretable and valuable. Although thoroughly documenting a data set is time-intensive, it results in clearer preservation and presentation of the data.

We sometimes use the term 'data' in a broader sense that encompasses other products that are generated alongside primary tabular and spatial data. These products fall into five general categories: raw data,

derived data, documentation, reports, and administrative records. To meet I&M Program goals, and to ensure adequate context for primary data products, these categories of products all require some level of management to ensure their quality and availability. We will use a more 'holistic view' about how natural resource data are generated, processed, finalized, and provided. All phases of data and information processing are integrated, and information about each phase and its processes must be shared through good documentation.

There are many potential sources of important data and information about the condition of natural resources in our parks. The types of work that may generate these natural resource data include:

- Inventories
- Monitoring
- Protocol development pilot studies
- Special focus studies done by internal staff, contractors, or cooperators
- External research projects
- Monitoring or research studies done by other agencies on park or adjacent lands
- Resource impact evaluations related to park planning and compliance with regulations
- Resource management and restoration work

Because the I&M Program focuses on long-term monitoring and natural resource inventories, our first priority is to produce and curate high-quality, well-documented data that we derive from these primary efforts. However, we can apply the same standards, procedures, infrastructure, and attitudes about data management to other natural resource data sources. As time and resources permit, we will work toward raising the level of data management for current projects, legacy data, and data originating from outside the I&M Program. We will place the greatest emphasis on those projects that are just beginning development and implementation because applying new data management practices to an ongoing project can be difficult and will generally meet with less success.

Goals and Objectives of Data Management

The data-related mission of the I&M Program is to provide scientifically and statistically sound data to support management decisions for the protection of park resources. The Program's success at identifying, cataloging, organizing, structuring, archiving, and disseminating relevant natural resource information will largely determine its effectiveness and standing among critics, peers, and advocates. The principal goal of the SODN Data Management Plan is to elucidate the driving concepts, principles, procedures, and processes for ensuring the quality, interpretability, security, longevity, and availability of ecological data and derived information produced by our inventory and monitoring efforts. Our objectives are centered around five main principles:

- *Quality* – ensure that appropriate quality assurance measures are taken during all phases of data development: acquisition, processing, summary and analysis, reporting, documenting, and archiving.
- *Interpretability* – ensure that complete documentation accompanies each data set so that users will be aware of its context, applicability, and limitations.
- *Security* – ensure that both digital and analog data are maintained and archived in a secure environment that provides appropriate levels of access to project leaders, technicians, network staff, and other users.
- *Longevity* – ensure that data sets are maintained in an accessible and interpretable format, accompanied by sufficient documentation.
- *Availability* – ensure that the data and information from our I&M studies are made available and easily accessible to managers and other users.

Data Stewardship Roles and Responsibilities

Everyone within the SODN I&M Program uses or manages data and information, and each of us has our roles and responsibilities in this process. This new and *crucial* emphasis on data management, analysis, and the reporting of results will require a large investment of personnel, time, and money, and the Network expects to invest at least thirty percent of available resources in developing and operating its data management system.

For the SODN I&M Program to work effectively, everyone within the Network will have stewardship responsibilities related to the production, analysis, management, and/or end use of the data.

The fundamental role of the Network data manager will be to coordinate these tasks. This requires understanding and determining program and project requirements, creating and maintaining data management infrastructure and standards, and communicating and working with all responsible individuals.

The data manager and the project leader are the personnel primarily responsible for data management. The network coordinator also assists by ensuring that project leaders meet timelines for data entry, verification, validation, summarization/analysis, and reporting. Figure 1 illustrates the core data management duties of the data manager and project leader and where those duties overlap.

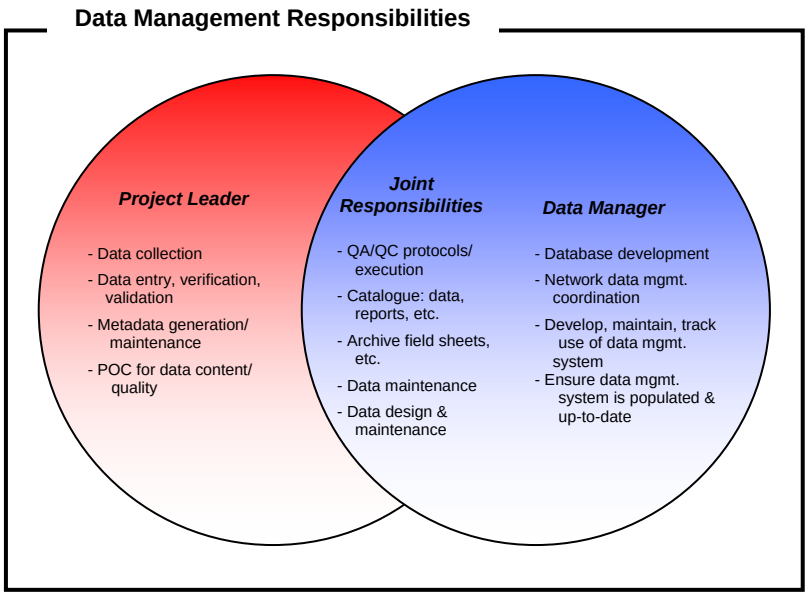

Figure 1. Core data stewardship responsibilities of project leaders and data managers.

Data Management Infrastructure/Architecture

A modern information management infrastructure (*e.g.*, staffing, hardware, software) represents the foundation upon which our network information system is built. Systems architecture refers to the applications, database systems, repositories, and software tools that make up the framework of our data management enterprise.

An important element of a data management system is a reliable, secure network of computers and servers maintained by national and local offsite IT specialists, assisted by network personnel. These individuals attend to hardware replacement, software installation and support, security updates, virus-protection, telecommunications networking, and server backups. Our digital infrastructure consists of two network data servers and servers maintained at the national level. Each of these components hosts different parts of our natural resource information system.

The national servers host and maintain online applications that provide storage and access to basic natural resource data and information collected by the I&M Program:

- *NatureBib* is the master database for natural resource-related bibliographic references.
- *NPSpecies* is the master database for species that occur in or near each park and the physical or written evidence for their occurrence (*e.g.*, vouchers, observations, references, and data sets).
- *NR-GIS Metadata Database* is the master metadata database for natural resource data sets.
- *NR-GIS Data Store* is a graphical search interface that links data set metadata to a searchable data server on which natural resource data sets are organized by NPS units, offices, and programs.

The network data servers host the following types of data and information:

- *Master project databases* – compiled data sets for monitoring projects and other multi-year efforts that have been certified for data quality.
- *Common lookup tables* – *e.g.*, parks, personnel, projects, species, etc.
- *Network digital library* – network repository for finished versions of products for Network projects (*e.g.*, reports, data set documentation, data files, formal metadata, etc.).
- *GIS files* – base spatial data, imagery, and project-specific themes.
- *Working files* – working databases, draft geospatial themes, draft reports, administrative records, etc.
- *Project tracking application* – database that tracks project status, contact information, product due dates.

Database Design Strategy

Rather than developing a single integrated database system, our approach uses modular, standalone project databases that share design standards and links to common lookup tables. Individual project databases are developed, maintained, and archived separately. There are several advantages to this strategy:

- Data sets are modular, allowing greater flexibility in accommodating the needs of each project area. Individual project databases and protocols can be developed at different rates without a significant effect on data integration. In addition, one project database can be modified without affecting the functionality of other project databases.
- By working up from modular data sets, we avoid a large initial investment in a centralized database and the concomitant difficulties of integrating among project areas with very different – and often unforeseen – structural requirements. Furthermore, the payoff for this initial investment is not always realized down the road by greater efficiency for interdisciplinary use.

Project database standards ensure compatibility among data sets, which is vital given the often unpredictable ways in which data sets will be aggregated and summarized. Well-thought-out standards also encourage sound database design and facilitate interpretability of data sets. The Network will follow the standards for database objects used by the Natural Resource Database Template (http://science.nature.nps.gov/im/apps/template/index.htm), to the extent possible. Databases that are developed for park and network projects will all contain the following main components:

- *Common lookup tables* that contain lists of parks, personnel, and species
- *Core tables and fields based on network and national templates* that contain 'who, where, and when' for project data collection
- *Project-specific fields and tables* containing recorded observations

Project Work Flow

From the perspective of managing workflow, there are two main types of projects:

- *Short-term projects*, which may include individual park research projects, inventories, or pilot studies done in preparation for long-term monitoring.
- *Long-term projects*, which will primarily include the implemented monitoring studies central to the I&M Program, but may also include multi-year research projects and monitoring performed by other agencies and cooperators. Long-term projects often require a higher level of documentation, review, and infrastructure development.

From a data management standpoint, a primary difference between short-term and long-term projects is an increased need to adhere to standards for long-term projects to ensure internal compatibility over time. This does not, however, negate the need to follow standards for short-term projects whenever possible and practical. Furthermore, both short- and long-term projects share many workflow characteristics, and both generate data products that must be managed and made available.

A project can be divided into five primary stages, each characterized by a particular set of activities carried out by staff involved with the project:

- *Planning and approval* – although this phase lacks specific data management activities, data managers must be kept informed of projects in this phase, particularly as timelines for products are finalized.
- *Design and testing* –collaboration between the project leader and the data manager is critical during this phase to assure data quality and integrity.
- *Implementation* – all aspects of this phase are overseen by the project leader; data management staff acts primarily as facilitators to support database applications, GIS, GPS, data validation, summarization, and analysis.
- *Product integration* – data management staff integrates products into national and network databases and repositories; data from working databases are merged into master databases.
- *Evaluation and closure* – the network coordinator, project leader, and data manager work together to assess how well the project met the stated objectives and what steps might be taken to make improvements.

During various phases of a project, the data take on different forms and are maintained in different places as they are acquired, processed, documented, and archived. Key points of this data life cycle are as follows:

- All raw data are archived intact.
- Working databases are the focal point of all modification, processing, and documentation of data collected for a given season (or other period that makes sense for a given project).
- Upon data certification, indicating that the data have passed all documentation and quality assurance requirements, the data are archived and posted or otherwise integrated with the national data applications.
- Data for long-term monitoring projects are uploaded into a master database that includes multiple years of data.
- Certified data sets are used to develop reports and other data products, which are also archived and posted to the appropriate national repositories.
- All subsequent revisions to certified data sets are documented in an edit log, which is distributed with the data.

Data Acquisition & Processing

The types of data handled by the I&M Program fall into three general classifications:

- *Programmatic data* are produced from projects that are either initiated (funded) by or substantially involve the I&M Program.
- *Non-programmatic NPS data* are produced by the NPS but did not involve the I&M Program.
- *Non-programmatic external data* are produced by agencies or institutions other than the NPS.

The importance or value placed on a data set in any of these categories will be based on its quality, completeness, relevance, and potential usefulness of the data set itself, as well as the impact it has on the SODN I&M Program and parks.

Most data acquired by the Network will be collected as field data (inventories and monitoring studies) or discovered through data mining initiatives (legacy/existing data). Methods of field data collection, such as paper field data forms, field computers, automated data loggers, and GPS units, will be specified in individual monitoring protocols and study plans. Field crew members will closely follow the established standard operating procedures (SOPs) in the project protocol.

The Network will conform to NPS standards and mandates, as well as to national I&M Program standards and procedures, to facilitate program integration and data/information sharing. General and protocol-specific SOPs will provide detailed instructions for processing specific types of data.

Quality Assurance/Quality Control

The perception that the data we collect during our inventory and monitoring studies are valuable resources to be used over the long-term is justified only if we have confidence in our data. Our efforts to detect trends and patterns in ecosystem processes require data of documented quality that minimize error and bias. Data of inconsistent or poor quality can result in loss of sensitivity and lead to incorrect interpretations and conclusions. We must remember that high quality data and information are vital to the credibility and success of the I&M Program, and everyone plays a part in ensuring that our products conform to these standards.

NPS Director's Order #11B: Ensuring Quality of Information Disseminated by the National Park Service specifies that information produced by the NPS must be of the highest quality and be based on reliable data sources that are accurate, timely, and representative of the most current information available. Therefore, we will establish and document procedures for quality assurance (QA) and quality control (QC) to identify and reduce the frequency and significance of errors at all stages in the data life cycle. When these procedures are followed, the progression from raw data to verified data to validated data implies increasing confidence in the quality of those data. The data manager will establish SOPs to ensure compliance with DO #11B. These procedures will document both internal and external review processes for data and information disseminated outside the network, as well as guidance for handling complaints about data quality.

Although many QA/QC procedures will depend upon the individual vital signs being monitored, some general concepts apply to all Network projects. Specific procedures to ensure data quality must be included in the protocols for each vital sign. Examples of QA/QC practices include:

- Standardized field data collection forms
- Use of field computers and automated data loggers
- Proper calibration and maintenance of equipment

- Field crew and data technician training
- Database features such as built-in pick lists and range limits to reduce data entry errors
- Automated error-checking routines

We evaluate data quality by applying verification and validation procedures. *Data verification* checks that the digitized data match the source data, and *data validation* checks that the data make sense. The Data Management Plan describes several methods for verifying and validating data, and each monitoring protocol will include specific procedures for assuring data quality.

A final report on data quality will be incorporated into the documentation for each project. This will include a listing of the specific methods used to assess data quality and an assessment of overall data quality prepared by the project leader.

Data Documentation

Data documentation is a critical step toward ensuring that all data sets retain their integrity and utility well into the future. Complete, thorough, and accurate documentation should be of the highest priority for long-term studies, and since long-term data sets are continually changing, this documentation must remain up-to-date. Data documentation refers to the development of metadata, which at the most basic level can be defined as 'data about data,' or more specifically as information about the content, context, structure, quality, and other characteristics of a data set. Additionally, standardized metadata provide a means to catalog data sets within intranet and internet systems, thus making these data sets available to a broad range of potential users.

Without metadata, potential users of a data set have little or no information regarding the quality, completeness, or manipulations performed on a particular 'copy' of a data set. Such ambiguity results in lost productivity as the user must invest time in tracking down information, or, worst case scenario, renders the data set useless because answers to these and other critical questions cannot be found. As such, data documentation must include an upfront investment in planning and organization.
At a minimum, we will require the following elements for documentation of all data managed by the Network:

- Data dictionaries and Entity Relationship Diagrams (ERDs) for all tabular databases
- Formal metadata compliant with Federal Geographic Data Committee (FGDC) standards, the National Biological Information Infrastructure (NBII) Profile (where appropriate), and the NPS Metadata Profile for all geospatial and biological data sets
- Project metadata

We will create all metadata according to NPS standards and guidelines. Formal metadata will be created using either Dataset Catalog, an NPS tool for producing abbreviated metadata, or the ArcCatalog data management application included with ArcGIS software, supplemented by the NPS Metadata Tools & Editor. We will publish metadata to the online NR-GIS Metadata Database, and all documentation will also be maintained with its accompanying data set(s) on the network data servers.

Support for Analysis & Reporting

Creating meaningful information from data sets through summaries and analyses is a critical component of the I&M Program and characterizes the Network's data management mission to provide useful information for park personnel. Close coordination between the project leader and data manager is important to identify opportunities and methods to streamline data extraction and exports from databases

based on project objectives, protocols, and data management and analysis SOPs. Where possible, project databases will include automated summary and report routines.

To make data sets available for subsequent analysis by third parties, the Network will establish a timeline of data processing steps including error-checking, summarizing, analyzing, and distributing data. Monitoring project leaders will be responsible for their project databases, but once a year they will review and certify the data set, write an annual report, and make the data available in a common repository for others to use in syntheses and further analyses.

Data Dissemination

One of the most important goals of the Inventory and Monitoring Program is to *integrate natural resource inventory and monitoring information into National Park Service planning, management, and decision-making.* To accomplish this goal, the Network will use a variety of distribution methods to make data and information collected and developed as part of the Program available to a wide community of users, including park staff, other researchers and scientists, and the public. We will ensure that:

- Data are easily discoverable and obtainable.
- Distributed data are accompanied by complete metadata that clearly establishes the data as a product of the NPS I&M Program.
- Data that have not yet been subjected to full quality control will not be released by the Network, unless necessary in response to a FOIA request *or* unless accompanied by a data quality disclaimer.
- Sensitive data are identified and protected from unauthorized access and inappropriate use.
- A complete record of data distribution/dissemination is maintained.

Distribution options include the Network data server and digital libraries, along with several online interfaces. The national I&M Program has developed several web-based applications and repositories to store different types of natural resource information:

- *NPSpecies* – data on park biodiversity (species information)
- *NatureBib* – park-related scientific citations
- *Biodiversity Data Store* – raw or manipulated data products that document the presence/absence, distribution, and/or abundance of any taxa in NPS units
- *NR-GIS Metadata and Data Store* – spatial and non-spatial metadata and accompanying data sets
- *Sonoran Desert Network website* – reports and metadata for all I&M data produced by the Network

Data Ownership

The NPS defines conditions for the ownership and sharing of collections, data, and results based on research funded by the United States government. All contracts and cooperative or interagency agreements should include clear provisions for data ownership and sharing as defined by the NPS:

- All data and materials collected or generated using NPS personnel and funds become the property of the NPS.
- Any important findings from research and educational activities should be promptly submitted for publication. Authorship must accurately reflect the contributions of those involved.
- Investigators must share collections, data, results, and supporting materials with other researchers whenever possible. In exceptional cases, where collections or data are sensitive or fragile, access may be limited.

As such, the Network has established guidelines for ensuring its ownership of data and other research information.

FOIA and Sensitive Data

The Freedom of Information Act (FOIA) stipulates that federal agencies, including the NPS, must provide access to agency records that are not protected from disclosure by exemptions. The NPS is directed to protect information about the nature and location of sensitive park resources under one Executive Order and four resource confidentiality laws:

- Executive Order No. 13007: Indian Sacred Sites
- National Parks Omnibus Management Act (NPOMA; 16 U.S.C. 5937)
- National Historic Preservation Act (16 U.S.C. 470w-3)
- Federal Cave Resources Protection Act (16 U.S.C. 4304)
- Archaeological Resources Protection Act (16 U.S.C. 470hh)

When any of these regulations are applicable, public access to data can be restricted. If disclosure could result in harm, information about the following natural resources may be classified as 'protected' or 'sensitive' and information withheld:

- Endangered, threatened, rare, or commercially valuable National Park System resources
- Mineral or paleontological sites
- Objects of cultural patrimony
- Significant caves

The Network will comply with all FOIA restrictions regarding the release of data and information, as instructed in NPS Director's Order #66 and accompanying Reference Manuals 66A and 66B (currently in draft). Classification of sensitive data will be the responsibility of Network staff, park superintendents, and project leaders. Network staff will classify sensitive data on a case-by-case, project-by-project basis and will work closely with project leaders to ensure that potentially sensitive park resources are identified, that information about these resources is tracked throughout the project, and that potentially sensitive information is removed from documents and products that will be released outside the Network.

Data Maintenance, Storage, and Archiving

Data, documents, and any other products that result from projects and activities that use Network data are all crucial pieces of information. Directions for managing these materials are provided in NPS Director's Order #19: Records Management (2001) and the accompanying NPS Records Disposition Schedule (NPS-19 Appendix B, revised 5-2003). This guidance states that records of natural and cultural resources are considered 'mission-critical' records (permanent records that are to be transferred to the National Archives when 30 years old) and that copies of these materials "should not, in any instance, be destroyed."

To ensure high-quality long-term management and maintenance of this information, the Network will implement procedures to protect information over time. These procedures will permit a broad range of users to easily obtain, share, and properly interpret both active and archived information, and they will ensure that digital and analog data and information are:

- Kept up-to-date in content and format so they remain easily accessible and usable

- Protected from catastrophic events (*e.g.*, fire and flood), user error, hardware failure, software failure or corruption, security breaches, and vandalism

Technological obsolescence is a significant cause of information loss, and data can quickly become inaccessible to users if they are stored in out-of-date software programs, on outmoded media, or on deteriorating (aging) media. Effective maintenance of digital files depends on the proper management of a continuously changing infrastructure of hardware, software, file formats, and storage media. Major changes in hardware can be expected every 1-2 years and in software every 1-5 years. As software and hardware evolve, data sets must be consistently migrated to new platforms or saved in formats that are independent of specific software or platforms (*e.g.*, ASCII delimited text files). Storage media should be refreshed (*i.e.*, copying data sets to new media) on a regular basis, depending upon the life expectancy of the media.

Regular backups of data and off-site storage of backup sets are the most important safeguards against data loss; therefore, we will establish data maintenance and backup schedules for data stored on the network data servers. Backups of data stored on personal workstations are the responsibility of each staff member. We strongly recommend that staff members store or regularly copy important files onto the network server. Backup routines represent a significant investment in hardware, media, and staff time; however, they are just a small percentage of the overall investment that we make in Program data.

Chapter 1. Introduction

> *Data and information are the basic products of scientific research. In ecological research, where field experiments and data collections can rarely be replicated under identical conditions, data represent a valuable and, often, irreplaceable resource . . . In long-term ecological studies, retention and documentation of high quality data are the foundation upon which the success of the overall project rests.*
>
> *Brunt 2000*

Information is the common currency among the many different activities and people involved in the stewardship of National Park Service (NPS) natural resources. As part of the Service's effort to "improve park management through greater reliance on scientific knowledge," a primary purpose of the Inventory and Monitoring (I&M) Program is to develop, organize, and make available natural resource data and to contribute to the Service's institutional knowledge. The I&M Program's ability to identify, catalog, organize, structure, archive, and disseminate relevant natural resource information will largely determine its efficacy and image among critics, peers, and advocates.

The NPS is a highly decentralized agency with complex data requirements. The primary audience for many of the products from the I&M Program is at the park level – park managers who need information to make better-informed decisions and to work more effectively with other agencies and individuals for the benefit of park resources. However, certain data are also needed at the regional or national level for a variety of purposes, and as stated by the National Park Advisory Board (2001), the findings "must be communicated to the public, for it is the broader public that will decide the fate of these resources."

To carry out this mission, the National Park Service initiated a service-wide natural resource Inventory and Monitoring Program encompassing 270 parks with significant natural resources. Ecologically similar parks were grouped into 32 networks. Each I&M Network has been tasked with 1) documenting existing park vertebrates and vascular plants (biological inventories) and 2) developing a management-based ecological monitoring program with a written plan and protocols. This includes a Data Management Plan that encompasses all aspects of the network program.

The Sonoran Desert Network (SODN) Data Management Plan presents the overarching strategy for ensuring that Program data are documented, secure, accessible, and useful for decades into the future. The plan also refers to other guidance documents, standard operating procedures (SOPs), and detailed protocols that convey more specific standards and steps for achieving our data management goals. The plan acts as a foundation upon which to build as new protocols are developed, advances in technology are adopted, and new concepts in data management philosophy are accepted.

This Plan describes how the Network will:

- Support I&M Program objectives
- Acquire and process data
- Assure data quality
- Document, analyze, and summarize data and information
- Integrate with nationally developed data management systems
- Disseminate data and information
- Maintain, store, and archive data

The SODN Data Management Plan covers I&M Program needs based on the most current information systems technology relevant through 2005. Revisions to this plan and associated data management

documents (guidelines and procedures) will be made as needed, and the overall plan will be reviewed and revised as necessary every 3-5 years.

1.1 Data and Data Management: An Overview

Collecting natural resource data is a critical step toward understanding the structure and function of the evolving ecosystems within our National Parks. We analyze and synthesize these 'raw' data to model various aspects of these ecosystems. In turn, we use our results and interpretations to make management decisions about the Park's vital natural resources. Thus, *data* collected by researchers and managed by the Sonoran Desert Network according to this Plan are transformed into *information* through analyses, syntheses, and modeling.

Any good set of data – whether collected last week or 20 years ago – must be accompanied by enough explanatory documentation (*e.g.*, how and why it was collected) so that we can understand it and use it with confidence. Therefore, our Network data management system cannot simply attend to the tables, fields, and values that make up a data set. It must also provide a process for developing, preserving, and integrating the context that makes data interpretable and valuable. Although thoroughly documenting a data set is time-intensive, it results in clearer preservation and presentation of the data.

We sometimes use the term 'data' in a broader sense that encompasses other products that are generated alongside primary tabular and spatial data. These products fall into five general categories: raw data, derived data, documentation, reports, and administrative records (Table 1.1).

Table 1.1. Categories of data and project products.

Category	Examples
Raw data	GPS rover files, raw field forms and notebooks, photographs and sound/video recordings, telemetry or remote-sensed data files, biological voucher specimens
Compiled/derived data	Relational databases, tabular data files, GIS layers, maps, species checklists
Documentation	Data collection protocols, data processing/analysis protocols, record of protocol changes, data dictionary, FGDC/NBII metadata, data design documentation, quality assurance report, catalog of specimens and photographs
Reports	Annual progress reports, final reports (technical or general audience), periodic trend analysis reports, publications
Administrative records	Contracts and agreements, study plans, research permits/applications, other critical administrative correspondence

To meet I&M Program goals and to ensure adequate context for primary data products, these categories of products all require some level of management to ensure their quality and availability. We will use a more 'holistic view' about how natural resource data are generated, processed, finalized, and provided. All phases of data and information processing are integrated, and information about each phase and its processes must be shared through good documentation.

1.1.1 Sources of Natural Resource Data

There are many potential sources of important data and information about the condition of natural resources in our parks. The types of work that may generate these natural resource data include:

- Inventories
- Monitoring
- Protocol development pilot studies
- Special focus studies done by internal staff, contractors, or cooperators
- External research projects
- Monitoring or research studies done by other agencies on park or adjacent lands
- Resource impact evaluations related to park planning and compliance with regulations
- Resource management and restoration work

Because the I&M Program focuses on long-term monitoring and natural resource inventories, our first priority is to produce and curate high-quality, well-documented data that we derive from these primary efforts. However, we can apply the same standards, procedures, infrastructure, and attitudes about data management to natural resource data acquired from other sources. As time and resources permit, we will work toward raising the level of data management for current projects, legacy data, and data originating from outside the I&M Program. We will place the greatest emphasis on those projects that are just beginning development and implementation because applying new data management practices to an ongoing project can be difficult and will generally meet with less success.

1.1.2 Types of Data Covered by the Data Management Plan

The Network coordinates or manages four major categories of data covered in this plan:

1) *Data managed in service-wide databases*. The Network uses three data systems developed by the I&M WASO office. *NatureBib* is used as a bibliographic tool for cataloging reports, publications, or other documents that relate to natural resources in park units. *Dataset Catalog* is used to document primarily non-spatial databases or other data assemblages. *NPSpecies* is used by the Network to develop and maintain lists of vertebrates and vascular plants in network parks, along with associated supporting evidence.
2) *Data developed or acquired directly by the Network through inventory, monitoring, or other projects*. This category includes project-related protocols, general protocols, reports, spatial data, and associated materials such as field notes and photographs provided to the Network by contractors or developed by network staff. Projects can be short-term (one to two years duration) or long-term (vital signs monitoring).
3) *Data that, while not developed or maintained by the Network, are used as data sources or provide context to other data sets.* Examples of this category include GIS data developed by parks or other agencies and organizations; national or international taxonomic or other classification systems; and air quality, climate, or hydrologic data collected by regional or national entities.
4) *Data acquired and maintained by network parks that the Network assists in managing.* Because of the lack of data management expertise in many parks, the Network provides data management assistance for high-priority data sets or those that may benefit from standardized procedures. Examples include data sets of both legacy and existing/ongoing natural resource monitoring data.

The above categories encompass one or more of the following data formats:

- Hard-copy documents (*e.g.*, reports, field notes, survey forms, maps, references, administrative documents)
- Objects (*e.g.*, specimens, samples, photographs, slides)
- Electronic files (*e.g.*, Word files, email, websites, digital images)
- Electronic tabular data (*e.g.*, databases, spreadsheets, tables, delimited files)
- Spatial data (*e.g.*, shapefiles, coverages, geodatabases, remote-sensing data)

1.2 Data Management Goals and Objectives

The long-term programmatic goals of the I&M Program are as follows:

- Establish natural resource inventory and monitoring standards throughout the National Park System that transcend traditional program, activity, and funding boundaries.
- Inventory the natural resources and park ecosystems under National Park Service stewardship.
- Integrate natural resource inventory and monitoring information into National Park Service planning, management, and decision making.
- Monitor park ecosystems to provide reference points for comparisons with other, altered environments.
- Share National Park Service accomplishments and information with other natural resource organizations and form partnerships for attaining common goals and objectives.

The data-related mission of the I&M Program is to provide scientifically and statistically sound data to support management decisions for the protection of park resources. The principal goal of the SODN Data Management Plan is to elucidate the driving concepts, principles, procedures, and processes for ensuring the quality, interpretability, security, longevity, and availability of ecological data and derived information produced by our inventory and monitoring efforts. Our objectives are centered around five main principles:

- *Quality* – ensure that appropriate quality assurance measures are taken during all phases of data development: acquisition, processing, summary and analysis, reporting, documenting, and archiving.
- *Interpretability* – ensure that complete documentation accompanies each data set so that users will be aware of its context, applicability, and limitations.
- *Security* – ensure that both digital and analog data are maintained and archived in a secure environment that provides appropriate levels of access to project leaders, technicians, network staff, and other users.
- *Longevity* – ensure that data sets are maintained in an accessible and interpretable format, accompanied by sufficient documentation.
- *Availability* – ensure that the data and information from our I&M studies are made available and easily accessible to managers and other users.

The SODN Data Management Plan outlines how we intend to implement and maintain a system that will serve the data and information management needs of the I&M Program. This plan reflects our commitment to the acquisition, maintenance, documentation, accessibility, and long-term availability of high-quality data and information.

1.3 Sonoran Desert Network Organization and Management

The Sonoran Desert Network encompasses 11 parks, ten in southern Arizona and one in southwestern New Mexico (Figure 1.1), ranging in size from 144 hectares (Tumacácori National Historical Park) to 133,882 hectares (Organ Pipe Cactus National Monument). With elevations from just over 300 meters to 2,600 meters, the parks contain a wide range of vegetation communities, from Sonoran desert scrub and mesquite woodlands at lower elevations to coniferous and montane riparian forests at higher elevations.

Figure 1.1. Parks in the Sonoran Desert Network.

Although the majority of parks in the Network were established to preserve cultural resources, all the units have significant natural resources. In fact, most of the cultural resource sites were chosen because of their proximity to indispensable natural resources, primarily water sources in the arid southwest. The management of cultural and natural resources is closely connected at these parks.

NPS staff for the Network consists of three permanent positions, a Network Coordinator, a Data Manager, and an Ecologist, and one term Cartographic Technician. Although not directly supported by the I&M Program, the network office also houses a regional Hydrologist who provides expertise to network parks. The network also has numerous research assistants working on various data management tasks, GIS projects, and monitoring protocols through a cooperative agreement with the Sonoran Institute. The Network works with the University of Arizona under another cooperative agreement that allows faculty, staff, and students to assist with projects when needed.

The network coordinator is accountable to the SODN Board of Directors (BOD) and is administratively supervised by the Chief of Natural Resources at the Southern Arizona Office (SOAR). The data manager, ecologist, and technician are supervised by the network coordinator. The research assistants are supervised by Sonoran Institute personnel.

The SODN I&M Program is overseen by a Board of Directors consisting of the nine superintendents of the network parks (two pairs of parks are operated as single administrative units with one superintendent) and the Chief of Natural Resources from the Southern Arizona Office. The BOD provides general direction to and oversight of I&M program activities, approves the budget and monitoring plan, and makes final decisions on the hiring of any new personnel.

The Technical Committee consists of a natural resources representative from each park administrative unit (including SOAR), an I&M Network representative, a chief of interpretation, and a cultural resources delegate. The committee members provide specific recommendations on technical issues and reviews of I&M plans and projects.

The Intermountain Region (IMR) Science Review Panel provides objective, expert reviews of network monitoring protocols, participates in network program reviews, and identifies opportunities for effective collaboration among IMR networks and other entities involved in ecological monitoring in the region.

1.4 Intended Audience and Layout for the SODN Data Management Plan

The conceptual model in Figure 1.2 shows the layout of the SODN Data Management Plan. There are three tiers of document types with multiple modules within tiers.

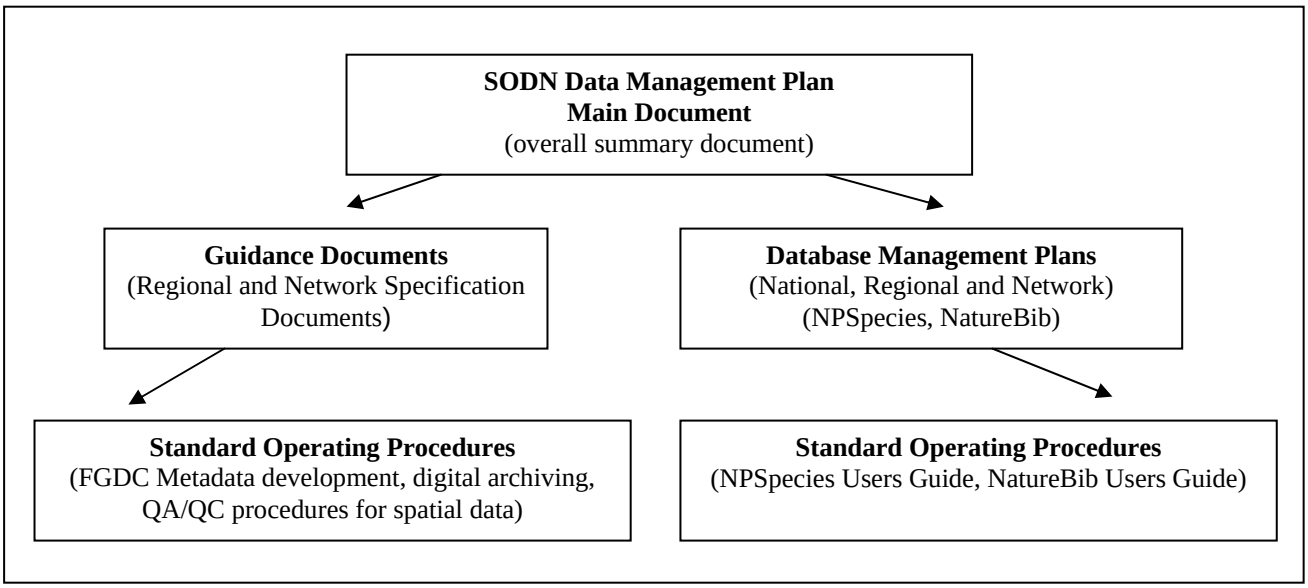

Figure 1.2. Layout of the SODN Data Management Plan.

We will develop four types of document as part of this plan:

1) The SODN Data Management Plan (summary document)
2) General Guidance Documents
3) Specific Database Management Plans
4) Standard Operating Procedures

We developed this data management plan in a modular format to increase both its readability and usability among a wider audience. This format allows plan users to easily locate and access subdocuments pertaining to a particular element of the data management process. Individual procedural documents from within the plan can be provided as standalone documents to park and regional data

management staff, cooperators, and project leaders. This format also allows for a more complete and technical review. Technical experts can be provided with applicable portions of the plan based on their expertise (GIS, databases, metadata, GPS, etc.).

Standard Operating Procedures (SOPs) developed as part of this plan detail work processes that are to be conducted or followed by network staff and cooperators. SOPs are useful because they thoroughly document the way activities are to be performed. The development and use of SOPs promotes data quality through consistent implementation of a process or procedure within the organization, even if there are temporary or permanent personnel changes.

Credits

Portions of the content of this chapter were adapted from material developed by Doug Wilder (Central Alaska Network) and Sara Stevens (Northeast Coastal and Barrier Network).

Chapter 2. Data Stewardship Roles and Responsibilities

Data management is about people and organizations as much as it is about information technology, database theories, and applications. Everyone within an organization uses or manages data and information. Thus, to serve the National Park Service and its constituents well, everyone in the Network must understand the flow of data and information, as well as our roles and responsibilities in this process. This new and *crucial* emphasis on data management, analysis, and the reporting of results requires a large investment of personnel, time, and money. The SODN I&M Program expects to invest at least thirty percent of available resources in developing and operating its data management program.

For the SODN I&M Program to work effectively, everyone within the Network will have stewardship responsibilities related to the production, analysis, management, and/or end use of the data. Table 2.1 summarizes the data stewardship roles of various network personnel. Each of these broad categories has principal, or 'must-do,' responsibilities, as well as many potential ancillary tasks. As coordinator of these tasks, the fundamental role of the Network Data Manager will be to understand and determine program and project requirements, to create and maintain data management infrastructure and standards, and to communicate and work with all responsible individuals.

Table 2.1. Categories of data stewardship involving all network personnel.

Stewardship Category	Related Activities	Principal Jobs or Positions
Note: Each position is listed in only one category according to overriding responsibilities. However, most positions contribute in each category.		
Production	Creating data or information from any original or derived source. This includes recording locations, images, measurements, and observations in the field; digitizing source maps; keying in data from a hardcopy source; converting existing data sources; image processing; and preparing and delivering informative products, such as summary tables, maps, charts, and reports.	Project Crew Member Project Crew Leader Data/GIS Specialist or Technician
Analysis	Using data to predict, qualify, and quantify ecosystem elements, structure, and function as part of the effort to understand these components, address monitoring objectives, and inform park and ecosystem management.	Network Ecologist Resource Specialist Statistician/Biometrician
Management	Preparing and executing policies, procedures, and activities that keep data and information resources organized, available, useful, compliant, and safe.	Network Data Manager Project Leader GIS Manager Information Technology Specialist Database Manager National Level I&M Data Manager Curator

End Use	Obtaining and applying available information to develop knowledge that contributes to understanding and managing park resources; informing the direction and scope of science information needs and activities.	Network Coordinator Park Managers and Superintendents Others

The remainder of this chapter discusses comprehensive data management roles and responsibilities that generally apply to all network activities. Each vital sign monitoring protocol and inventory study plan contains specific instructions for assignments and tasks that nest within this overall framework. Individuals who carry out monitoring protocols and inventory study plans are responsible for reading and understanding these instructional guidelines.

2.1 Roles and Responsibilities

A *Role* is a function or position (*e.g., Data Manager*).
A *Responsibility* is a duty or obligation (*e.g., review data records*).

We cannot and do not expect any individual to do everything in the data management process. No one person can work effectively as data producer, analyst, manager, and end user. While this was often the case in the past (and often still is), we can no longer follow this model. The current and expected capacity, diversity, and rate of change in information technology make managing any large amount of information a greater task than any individual can expect to do alone.

An increasing demand for more detailed, higher quality data and information about natural resources and ecosystem functions requires a group of people working together to steward data and information assets. Knowledgeable individuals from many areas must come together to ensure that data are collected using appropriate methods and that resulting data sets, reports, maps, models, and other derived products are well managed. Data sets and the presentations of these data must be credible, representative, and available for current and future needs.

Table 2.2 summarizes the roles and responsibilities of various personnel (see Appendix 2.1 for more complete descriptions of roles). These roles are listed 'from the ground up' to help demonstrate the hierarchy and overlap of responsibilities. For example, a project leader is ultimately responsible for the activities listed in the field level roles of crew leader and crew member. In addition, the network coordinator ensures that the network data manager and ecologist achieve the required performance level.

Table 2.2. Summary of roles and responsibilities.

Role	Primary Responsibilities Related to Data Management
Project Crew Member	Record and verify measurements and observations based on project objectives and protocols. Document methods, procedures, and anomalies.
Project Crew Leader	Supervise Crew Members to ensure their data collection and management obligations are met, including data verification and documentation.
Data/GIS Specialist or Technician	Perform assigned level of technical data management and/or GIS activities, including data entry, data conversion, and documentation. Work on overall data quality and stewardship with Project Leaders, Resource Specialists, and the Network Data Manager.

Role	Primary Responsibilities Related to Data Management
Information Technology/Systems Specialist	Provide and maintain an information systems and technology foundation to support data management.
Project Leader	Oversee and direct operations, including data management requirements, for one or more network projects. Maintain communication with Project Staff, Network Data Manager, and Resource Specialist regarding data management. **Note:** The Project Leader is often a Resource Specialist, in which case the associated responsibilities for data authority apply (see next role). A Project Leader without the required background to act as an authority for the data will consult with and involve the appropriate Resource Specialists.
Resource Specialist	Understand the objectives of the project, the resulting data, and their scientific and management relevance. Make decisions about data with regard to validity, utility, sensitivity, and availability. Describe, publish, release, and discuss the data and associated information products. **Note:** The Resource Specialist serving as a Project Leader is also responsible for the duties listed with that role.
GIS Manager	Support network management objectives. Coordinate and integrate local GIS and resource information management with Network, Regional, and National standards and guidelines.
Network Data Manager	Provide overall network planning, training, and operational support for the awareness, coordination, and integration of data and information management activities, including people, information needs, data, software, and hardware. Serve as Point-of-Contact for National Park Service database applications (NPSpecies, NatureBib). Coordinate internal and external data management activities.
Curator	Oversee all aspects of the acquisition, documentation, preservation, and use of park collections.
Statistician or Biometrician	Analyze data and present information according to established protocols.
Database Manager	Apply particular knowledge and abilities related to database software and associated application(s).
Network Ecologist	Ensure useful data are collected and managed by integrating natural resource science in network activities and products, including objective setting, sample design, data analysis, synthesis, and reporting.
Network Coordinator	Ensure programmatic data and information management requirements are met as part of overall network business.

Role	Primary Responsibilities Related to Data Management
I&M Data Manager (National Level)	Provide service-wide database design, support, and services, including receiving and processing to convert, store, and archive data in service-wide databases.
Other End Users (Managers, Superintendents, Scientists, Public)	Ensure data and derived products are used and applied appropriately. Provide feedback for improvements in data and products.

2.2 Data Management Coordination

The Natural Resource Challenge states that collaboration among the National Park Service, other public agencies, universities, and non-governmental organizations is necessary to effectively acquire, apply, and promulgate the scientific knowledge gained in National Parks. The Inventory and Monitoring Program encourages coordination among participants at all levels to help ensure that data collected by NPS staff, cooperators, researchers, and others are entered, quality-checked, documented, analyzed, reported, archived, cataloged, and made available for management decision-making, research, and education.

The network data manager works with national I&M Program data management staff and regional resource information management personnel to maintain a high level of involvement in service-wide and regional databases and data management policy. The network data manager works locally with network personnel, park staff, and cooperators to promote and develop workable standards and procedures that result in the integration and availability of data sets.

Key contacts for the network data manager include park GIS and data managers and the project leaders for each monitoring or inventory project. Consistent communication among these personnel leads to common understanding and better synchronization of network and park data management activities. Park and network staff work together to manage resource information using a variety of methods. These include personal visits, phone calls, email, joint meetings and training sessions, as well as the meetings and work of the Network's Technical Committee and Board of Directors.

Involvement and input from park scientists and resource information management staff is essential. We rely on everyone within the Network for the successful development of planning materials, inventory study plans, and monitoring protocols.

2.2.1 Network of Networks

Data managers throughout the Program regularly coordinate with each other and national program staff via annual meetings, conference calls, workgroups, a listserv, websites, and informal communication. Data managers from these networks share the workload by collaborating to develop their respective Network Data Management Plans. This model of cooperation and communication is effective, and it can be applied both to resource information management and administrative issues.

The Sonoran Desert Network maintains an active role in promoting practical consistency among protocols and data sets involving other networks and organizations. Staff from all network parks is encouraged to participate in program development and activities, to use Inventory and Monitoring Program resources, and to communicate with Network and Program staff to share information about progress, direction, and concerns.

2.3 Project Stewardship

Since the data management aspects of every inventory or monitoring project normally require the expertise and involvement of several people over a period of months or years, it makes sense that one person is charged with keeping track of the objectives, requirements, and progress for each project. This project leader (or *steward*) is usually a resource management specialist with training and experience in the field of science related to the inventory or monitoring project.

The project leader may have worked in the geographic area where the study occurs. This can be helpful as it gives the project leader a background of information to use when overseeing fieldwork, coordinating with GIS and data managers on information management needs, evaluating data resources, and understanding the project's objectives. A project leader who cannot act as an authority for the data should work with the appropriate resource specialists to account for those aspects of data stewardship.

To ensure the quality of each project, including data requirements, project leaders should be assigned only those projects they can effectively oversee based on workload and other relevant factors. Unless the project is short-term (three months or less), the project will have at least one alternate or backup project leader to provide continuity in case the principal project leader becomes temporarily or completely unavailable.

2.3.1 Data Stewardship – Sharing Responsibilities

Keeping track of data from the time of acquisition until they are no longer useful is the shared responsibility of everyone involved with data – producers, analysts, managers, and end users. This, in essence, is *data stewardship*. It is a principle of mutual accountability rather than one particular job for one individual.

Successful data stewardship requires that people involved in Network activities learn and understand network expectations for continuous data management. This is equally important for network staff, park employees, and contractors or cooperators. All project participants receive training, briefings, materials, and additional regular communication about data stewardship from supervisors, project leaders, and data managers. The purpose is to promote the appropriate level of understanding about how their efforts relate to park and network management objectives, National Park Service and Department of Interior policies, and other federal government requirements. Other relevant context and linkages can also be discussed to help establish a sense of ownership and accountability within project staff.

Inventory and Monitoring project leaders must understand resource information management issues and requirements, and they must be aware of the challenges and limitations of field data collection, including the use of GPS. We can achieve these goals through training, through detailed and regular briefings, and by providing well-trained field crews to collect data at reasonable intervals.

2.3.2 Documentation is Paramount

If one shared responsibility stands out in importance and value, it is the careful documentation of data sets, the data source(s), and the methodology by which the data were collected or acquired. This documentation establishes the basis for the appropriate use of the data in resulting analyses and products, both in the short and long term. Network monitoring protocols specify key elements of data documentation, and any network data collected according to these protocols will include the name, date, and version of the associated protocol. (See Chapter 7 of this plan for important guidance on documentation and metadata.)

2.3.3 The Hub of Data Stewardship

Project leaders, data managers, and GIS managers form the central data management team for inventory and monitoring projects. Each is responsible for certain aspects of project data, and all share responsibility for some overlapping tasks. Because of the collaborative nature of project data management, good communication among project leaders, data managers, and GIS managers is essential to meeting program goals. Figure 2.1 illustrates some of the overlap between project data management responsibilities.

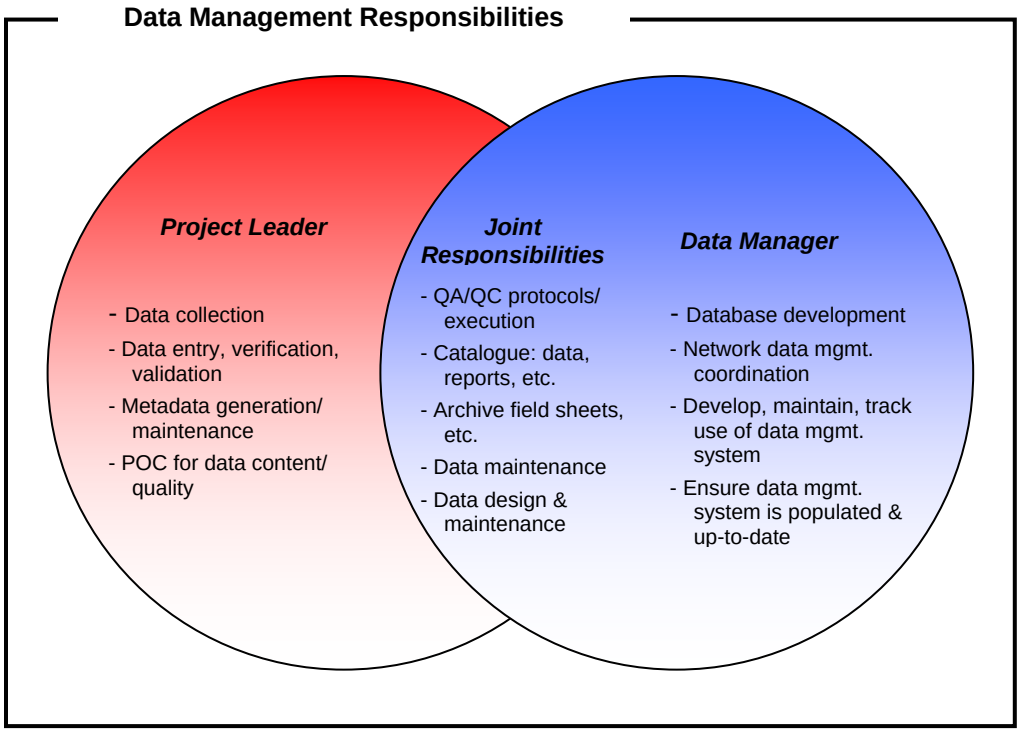

Figure 2.1. Shared responsibilities for each project.

2.3.3.1 Project Leaders (Stewards)

The project leader is responsible for data quality during all phases of the project, including collecting, entering, handling, reviewing, summarizing, and reporting data. Developing project documentation and metadata are crucial elements of the project leader's role.

Specifically, a project leader is responsible for:

- *Project documentation* that describes the 'who, what, where, when, why, and how' of a project
- Documentation and implementation of *standard procedures* for field data collection and data handling
- *Quality assurance and quality control measures*, which include the supervision and certification of all field operations, staff training, equipment calibration, species identification, data collection, data entry, verification, and validation
- Maintenance of concise explanatory documentation of all deviations from standard procedures
- Detailed documentation for each field data collection period
- Maintenance of hard copies of data forms and archiving of original data forms

- Scheduling of regular project milestones such as data collection periods, data processing target dates, and reporting deadlines
- Regular summary reports, periodic trend analyses of data, resulting reports, and their public availability
- Acting as the main point of contact concerning data content

The project leader will also work closely with the data manager to:

- Develop quality assurance and quality control procedures specific to project operations.
- Identify training needs for staff related to data management philosophy, database software use, and quality control procedures.
- Coordinate changes to the field data forms and the user interface for the project database.
- Document and maintain master data.
- Identify sensitive information that requires special consideration prior to distribution.
- Manage the archival process to ensure regular archival of project documentation, original field data, databases, reports and summaries, and other products from the project.
- Define the process of how project data will be transformed from raw data into meaningful information.
- Create data summary procedures to automate and standardize this transformation process.
- Identify and prioritize legacy data for conversion and convert priority data sets to a modern format.
- Increase the interpretability and accessibility of existing natural resource information.

2.3.3.2 Data Managers

The data manager is responsible for ensuring the compatibility of project data with program standards, for designing the infrastructure for the project data, and for ensuring long-term data integrity, security, and availability.

Data managers will:

- Develop and maintain the infrastructure for metadata creation, project documentation, and project data management.
- Create and maintain project databases in accordance with best practices and current program standards.
- Provide training in the theory and practice of data management tailored to the needs of project personnel.
- Develop ways to improve the accessibility and transparency of digital data.
- Establish and implement procedures to protect sensitive data according to project needs.
- Collaborate with GIS specialists to integrate tabular data with geospatial data in a GIS system in a manner that meets project objectives.

Data managers will also work closely with the project leader to:

- Define the scope of the project data and create a data structure that meets project needs.
- Become familiar with how the data are collected, handled, and used.
- Review quality control and quality assurance aspects of project protocols and standard procedure documentation.
- Identify elements that can be built into the database structure to facilitate quality control, such as required fields, range limits, pick-lists, and conditional validation rules.

- Create a user interface that streamlines the process of data entry, review, validation, and summarization that is consistent with the capabilities of the project staff.
- Develop automated database procedures to improve the efficiency of the data summarization and reporting process.
- Make sure that project documentation is complete, complies with metadata requirements, and enhances the interpretability and longevity of the project data.
- Ensure regular archival of project materials.
- Inform project staff of changes and advances in data management practices.

2.3.3.3 GIS Managers

The GIS manager administers spatial data themes associated with SODN I&M projects, as well as other spatial data related to the full range of park resources. They incorporate spatial data into the GIS. They also maintain standards for geographic data and are responsible for sharing and disseminating GIS data throughout the Network.

The GIS manager works in collaboration with project leaders to:

- Determine the GIS data and analysis needs for the project.
- Develop procedures for field collection of spatial data including the use of GPS and other spatial data collection techniques.
- Display, analyze, and create maps from spatial data to meet project objectives.
- Properly document data in compliance with spatial metadata standards.

GIS managers will also work directly with data managers to:

- Design databases and other applications for the Network.
- Create relationships between GIS and non-spatial data.
- Create database and GIS applications to facilitate the integration and analysis of both spatial and non-spatial data.
- Establish and implement procedures to protect sensitive spatial data according to project needs.
- Develop and maintain an infrastructure for metadata creation and maintenance.
- Ensure that project metadata are created and comply with national and agency standards.

Credits

This chapter was adapted from material developed by Rob Daley (Greater Yellowstone Network) and the NPS Prairie Cluster Prototype Data Management Plan (2002).

Chapter 3. Data Management Resources: Infrastructure and Systems Architecture

Our computer resource infrastructure is composed of computers and servers that are functionally or directly linked through computer networking services. This infrastructure represents the foundation upon which our network information system is built. Systems architecture signifies the applications, database systems, repositories, and software tools that make up the framework of our data management enterprise.

The SODN I&M Program relies on local off-site, regional, and national IT personnel and resources, as well as network staff, to maintain its computer resource infrastructure. This includes but is not limited to hardware replacement, software installation and support, security updates, virus-protection, telecommunications networking, and backups of servers. Communication with park and regional IT specialists is essential to ensure adequate resources and service continuity for systems architecture. Rather than focusing on a detailed description of current computer resources, this chapter will instead describe the network infrastructure in more general terms and focus more specifically on the systems architecture that is central to data management.

3.1 Computer Resources Infrastructure

An important element of a data management program is a reliable, secure network of computers and servers. Our digital infrastructure has two main components: network data servers and servers maintained at the national level (Figure 3.1). The primary responsibility for maintaining this infrastructure lies with the Southern Arizona Office Computer Specialist in Phoenix, who may be assisted by two local IT specialists at Saguaro National Park and the Western Archeological and Conservation Center. Some basic IT functions are performed by network personnel. Working together, we administer all aspects of system security and backups.

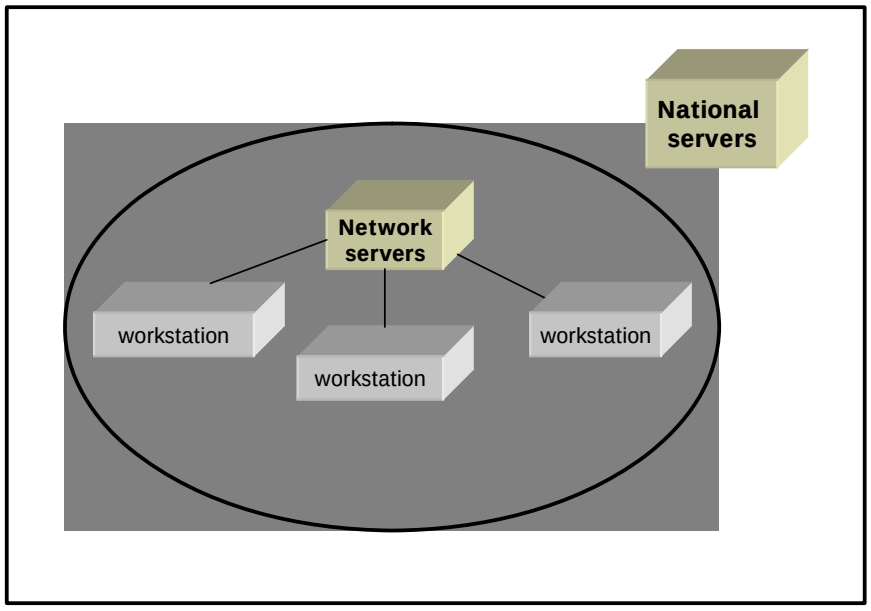

Figure 3.1. Schematic representing the logical layout and connectivity of computer resources.

These national and network components each host different parts of our natural resource information system.

National servers

- Master applications – integrated client-server versions of NatureBib, NPSpecies, NR-GIS Metadata Database
- Centralized repositories – NR-GIS Data Store, Protocol Clearinghouse, Biodiversity Data Store
- Public access sites – portals to NatureBib, NPSpecies, NPSFocus, STORET, and websites for monitoring networks

Network data servers

- Master project databases – compiled data sets for monitoring projects and other multi-year efforts that have been certified for data quality
- Common lookup tables – *e.g.*, parks, projects, personnel, species
- Project tracking application – used to track project status, contact information, product due dates
- Network digital library – repository for finished versions of products from Network projects (*e.g.*, reports, methods documentation, data files, metadata, etc.)
- GIS files – base spatial data, imagery, project-specific themes
- Working files – working databases, draft geospatial themes, drafts of reports, administrative records

3.1.1 Maintaining Digital Files

The SODN local area network (LAN) accommodates hierarchical directory structures for storing digital files. There are six main categories of directory structure sections in which digital files can be maintained:

1) *Admin* – documents related to program administration.
2) *Databases* – local copies of national databases (*e.g.*, NPSpecies), databases for common lookup tables, and back-end databases for the SODN Library and Dataset Catalog.
3) *Libraries* – read-only storage of cataloged photographs and other reference documents generated and maintained by the Program.
4) *Working* – workspace where groups and individuals can maintain draft material and other files as arranged by projects. The layout of folders and subfolders is more flexible here than elsewhere. On an annual basis, project leaders and administrators should work to clean out these sections by identifying material that belongs in one of the Libraries or the Project Archive, deleting unneeded files, and moving the remainder to the appropriate location.
5) *GIS* – base spatial data, imagery, and project-specific themes. The typical arrangement of folders and subfolders is park-specific. Some project-specific themes in development reside in working sections until they are migrated here upon being finalized. This section is housed on the archive server due to storage space requirements and backup issues.
6) *Project Archive* – read-only storage of finished project products.

The key aspects of this file management strategy are as follows:

- Working files are kept separate from finished products.
- Finished products are typically read-only, except for 'inbox' folders where users can drop things off to be cataloged and filed.
- Standards such as naming conventions and hierarchical filing are enforced within the Libraries, Project Archive, Database, and GIS sections. Although less stringent in other sections, these conventions are encouraged as good practice.

3.2 National Information Management Systems

The need for effective natural resource information management cuts across NPS divisional boundaries, and management strategies must be defined at the highest level possible. In this context, integrated inventory and monitoring of natural resources is multidisciplinary and requires national-level programmatic data and information management strategies for success.

The basic strategy of natural resource, and therefore inventory and monitoring, information management is to provide integrated natural resource databases and information systems that enhance NPS managers' and staff's access and use of timely and valid data and information for management decisions, resource protection, and interpretation. Inventory and monitoring information needs are broadly separated into two categories:

- *Detailed data and information needed for onsite resource management and protection*. The information used to guide natural resource management decisions must be specific to inform and be useful to management staff at parks and central offices.
- *Summary information needed to describe the resources and their condition*. This kind of information is usually aggregated across the National Park Service for use by NPS and DOI managers and central office personnel to answer requests from Congress and for budget, program, and project planning.

The NPS Natural Resource Program Center (NRPC) and the I&M Program actively develop and implement a national-level, program-wide information management framework. NRPC and I&M staff integrate desktop database applications with Internet-based databases to serve both local and national-level data and information requirements. NRPC staff members work with regional and support office staff to develop extensible desktop GIS systems that integrate closely with the database systems. Centralized data archiving and distribution capabilities at the NRPC provide for long-term data security and storage. NRPC sponsors training courses on data management, I&M techniques, and remote sensing to assist I&M data managers with developing and effectively utilizing natural resource information.

3.2.1 Core National-Level Application Architecture

To achieve an integrated information management system, three of the national-level data management applications (NatureBib, NPSpecies, and NR-GIS Metadata Database) utilize a distributed application architecture with both desktop and Internet-accessible (master) components (Figure 3.2).

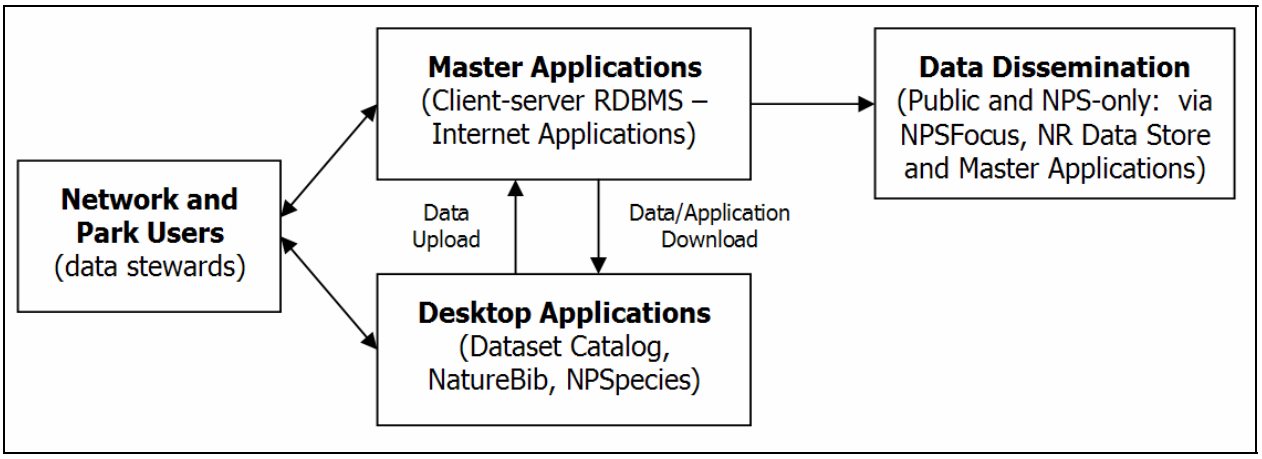

Figure 3.2. Model of the national-level application architecture.

NatureBib

NatureBib is the master database for bibliographic references that merges a number of previously separate databases such as Whitetail Deer Management Bibliography (DeerBib), Geologic Resource Bibliography (GRBib), and others. It also contains citation data from independent databases such as NPSpecies, the Dataset Catalog, and the NR-GIS Metadata Database. It currently focuses on natural resource references, but may eventually link to references on cultural resources and other park operations. As with NPSpecies and the NR-GIS Metadata Database, it is possible to download data from the master web version into the MS Access desktop version that can be used locally on computers with limited Internet connectivity *(http://www.nature.nps.gov/nrbib)*.

NPSpecies

NPSpecies is the master species database for the NPS. The database lists the species that occur in or near each park and the physical or written evidence for the occurrence of the species (*e.g.*, references, data sets, vouchers, and observations). Taxonomy and nomenclature are based on ITIS, the interagency Integrated Taxonomic Information System. The master version of NPSpecies for each park or network can be downloaded from the master website into an MS Access version of NPSpecies. The Internet-based version is the master database, which can be accessed via password-protected logins administered by park, network, and regional data stewards (Points-of-Contact) assigned for each park and network. The master database requires that species lists be certified by networks before any data will be made available to the public. NPSpecies is linked to NatureBib for bibliographic references that provide written evidence of a species' occurrence in a park and will be linked to the NR-GIS Metadata Database to document biological inventory products. The MS Access application and additional details can be found at the NPSpecies website *(http://science.nature.nps.gov/im/apps/npspp/index.htm)*.

Dataset Catalog and the NR-GIS Metadata Database

Dataset Catalog is a desktop metadata database application developed by the I&M Program to provide a tool that parks, networks, and cooperators can use to inventory and manage data set holdings. Although not designed as a comprehensive metadata tool, the Dataset Catalog is used for cataloging abbreviated metadata about a variety of digital and non-digital natural resource data sets. The Dataset Catalog helps parks and networks begin to meet Executive Order 12906 mandating federal agencies to document all data collected after January 1995. It provides brief metadata and a comprehensive list about all resource data sets for use in data management, project planning, and more stringent metadata activities. As with other service-wide applications, the master metadata database (NR-GIS Metadata Database) is available through a website and will be linked to NPSpecies and NatureBib. It will be possible to download a

version in MS Access format from the master website (*Dataset Catalog: http://science.nature.nps.gov/im/apps/datacat/index.htm and NR-GIS Metadata Database: http://science.nature.nps.gov/nrdata).*

3.2.2 Other National-Level I&M Information Management and GIS Applications

NPSTORET

STORET is an interagency water quality database developed and supported by the Environmental Protection Agency (EPA) to house local, state, and federal water quality data collected in support of managing the nation's water resources under the Clean Water Act. STORET is used by the NPS as a repository of physical, chemical, biological, and other monitoring data collected in and around national park units by park staff, contractors, and cooperators. The NPS operates its own service-wide copy of STORET and makes periodic uploads to the EPA STORET National Data Warehouse so that data collected by and for parks will be accessible to the public. NPS Director's Order (DO) #77 indicates that the NPS should archive water quality data in STORET, and the NPS Water Resources Division (WRD) requires that any data collected as part of a funded WRD project be archived in STORET. NPSTORET (also known as Water Quality Database Templates) is the NPS master database designed to facilitate park-level standardized reporting for STORET. Metadata, protocols, data dictionaries, and reporting capabilities are available through a front-end form, and network staff and cooperators can use the MS Access version of NPSTORET either as a direct database for data entry and management or as a means of submitting data for upload to STORET by WRD staff. The MS Access application and additional details can be found at *http://www.nature.nps.gov/water/infoanddata/index.htm*. Additional information on STORET can be found at *http://www.epa.gov/storet*.

Natural Resource Database Template

The Natural Resource Database Template (NRDT) is a flexible, relational database in MS Access for storing inventory and monitoring data (including raw data collected during field studies). This relational database can be used as a standalone database or in conjunction with GIS software (*e.g.*, ArcView or ArcGIS) to enter, store, retrieve, and otherwise manage natural resource information. The template has a core database structure that can be modified and extended by different parks and networks depending on the components of their inventory and monitoring program and the specific sampling protocols they use. The Natural Resource Database Template is a key component of the I&M Program's standardized monitoring protocols. These monitoring protocols include separate modules detailing different aspects of monitoring project implementation, from sampling design to data analysis and reporting, and include data management components that describe database table structure, data entry forms, and quality checking routines. Approved monitoring protocols, including the databases that are based on the Database Template, are made available through a web-based protocol clearinghouse (see below). A description of the Database Template application, a data dictionary, and example implementations are located on the NR Database Template website *(http://science.nature.nps.gov/im/apps/template/index.htm).*

Natural Resource Monitoring Protocols Clearinghouse

The Natural Resource Monitoring Protocol Clearinghouse (*i.e.*, Protocol Database) is a web-based clearinghouse of sampling protocols used in national parks to monitor the condition of selected natural resources. The database provides a summary of, and in many cases allows the user to download a digital copy of, sampling protocols that have been developed by the prototype monitoring parks or other well-established protocols used in National Parks. The Protocol Database also makes it possible to download database components (*e.g.*, tables, queries, data entry forms) consistent with the Natural Resource Database Template that have been developed for a particular protocol in MS Access. See the Protocol Database website for available protocols (*http://science.nature.nps.gov/im/monitor/protocoldb.cfm*).

NR-GIS Data Store

The NR-GIS Data Store is a key component of the data dissemination strategy employed by the I&M Program. The NR-GIS Data Store is a graphical search interface that links data set metadata to a searchable data server on which data sets are organized by NPS units, offices, and programs. The interface allows customized public or protected searches of natural resource data sets, inventory products, and GIS data produced by the I&M and Natural Resource GIS Programs. Each park or network is able to post and curate its data on the server. The NR-GIS Data Store will be integrated with the master NR-GIS Metadata Database application to streamline programmatic data documentation and dissemination processes. The simple browse function of this server can be accessed at: *http://nrdata.nps.gov/*. See the NR-GIS Data Store website for further information (*http://science.nature.nps.gov/nrdata*).

3.3 Network Systems Architecture

Rather than developing a single, integrated database system, our data design consists of modular, standalone project databases that share design standards and links to centralized data tables. Individual project databases are developed, maintained, and archived separately. There are numerous advantages to this strategy:

- Data sets are modular, allowing greater flexibility in accommodating the needs of each project area. Individual project databases and protocols can be developed at different rates without a significant cost to data integration. In addition, one project database can be modified without affecting the functionality of other project databases.
- By working up from modular data sets, we avoid a large initial investment in a centralized database and the concomitant difficulties of integrating among project areas with very different – and often unforeseen – structural requirements. Furthermore, the payoff for this initial investment may not be realized down the road by greater efficiency for interdisciplinary use.

3.3.1 Project Database Standards

Project database standards are necessary for ensuring compatibility among data sets, which is vital given the often unpredictable ways in which data sets will be aggregated and summarized. When well thought out, standards also help to encourage sound database design and facilitate interpretability of data sets. As much as possible, network standards for fields, tables, and other database objects will mirror those conveyed through the Natural Resource Database Template. Where there are differences between local and national standards, documentation of the rationale for these differences will be developed. In addition, documentation and database tools (*e.g.*, queries that rename or reformat data) will be developed to ensure that data exports for integration are in a format compatible with current national standards. Databases that are developed for park and network projects will all contain the following main components:

- *Common lookup tables* – links to entire tables that reside in a centralized database, rather than storing redundant information in each database. These tables typically contain information that is not project-specific (*e.g.*, lists of parks, personnel, and species).
- *Core tables and fields based on network and national templates* – these tables and fields are used to manage the information describing the 'who, where and when' of project data. Core tables are distinguished from common lookup tables in that they reside in each individual project database and are populated locally. These core tables contain critical data fields that are standardized with regard to data types, field names, and domain ranges.
- *Project-specific fields and tables* – the remainder of database objects can be considered project-specific, although there will typically be a large amount of overlap among projects. This is true

even among projects that may not seem logically related – for example, a temperature field will require similar data types and domain values. As much as is possible, efforts will be made to develop these project-specific objects to be compatible with those maintained by other networks and cooperators managing similar data sets, especially if integration with other data sets is important for meeting project objectives.

3.3.1.1 Centralized Database Components: Common Tables

Certain key information is not only common to multiple data sets, but to the organization as a whole – lists of contacts, projects, parks, and species that are often complex and dynamic. It is a good strategy to centralize this information so that users have access to the most updated versions in a single, known place. Centralizing also avoids redundancy and versioning issues among multiple copies. Centralized information is maintained in database tables that can be linked or referred to from several distinct project databases (Figure 3.3). Network applications for project tracking, administrative reporting, or budget management can also link to the same tables so that all users in the Network have instantaneous access to edits made by other users.

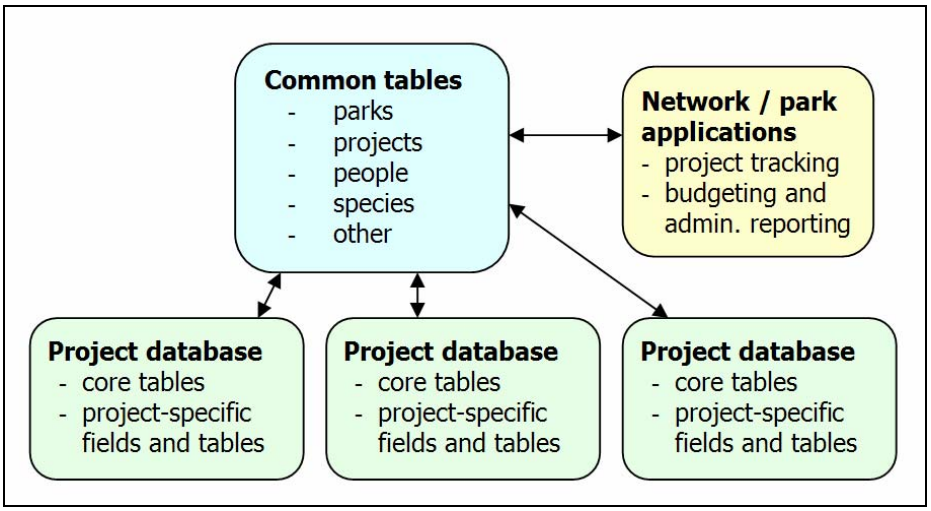

Figure 3.3. Common lookup tables and satellite databases.

At present, these common tables are grouped and maintained in separate MS Access (.mdb) files as shown in Table 3.1. Separating these tables by functional groupings is done primarily to reduce conflicts and performance losses associated with multiple users in MS Access. Databases associated with individual projects each access the common tables via links established in each project back-end data file. These components reside on the network server to provide access to individual project databases.

Table 3.1. Groupings for common lookup tables.

Grouping	Description
Parks	List of park units and networks
Projects	List of park and network projects, including inventories, monitoring, park-sponsored initiative projects, and external research projects
People	Comprehensive list of contacts for parks and Network, project-specific crew lists, lists of groups and users for tracking and managing access privileges

Species	Comprehensive list of taxa for the network parks, linkage to NPSpecies taxonomic module, project-specific species lists
Other Lookups	Lists of watersheds, drainages, place names, weather conditions, habitat attributes

3.3.1.2 Different Levels of Data Standards

The three types of database objects also correspond to three acknowledged levels of data standards. Because common lookup tables are stored in one place and are referred to by multiple databases, they represent the highest level of data standard because they are implemented identically among data sets. The second level of standards is implied by the core template fields and tables, which are standardized where possible, but project-specific objectives and needs could lead to varied implementations among projects. The third level of standards is applied most flexibly to accommodate the range of needs and possibilities for each project, yet always with compatibility and integrity in mind. The following figure (3.4) presents the resulting variation in implementation of these differing levels as a 'bull's eye,' with the common lookup tables providing the most consistent implementation and hence the smallest range of variation.

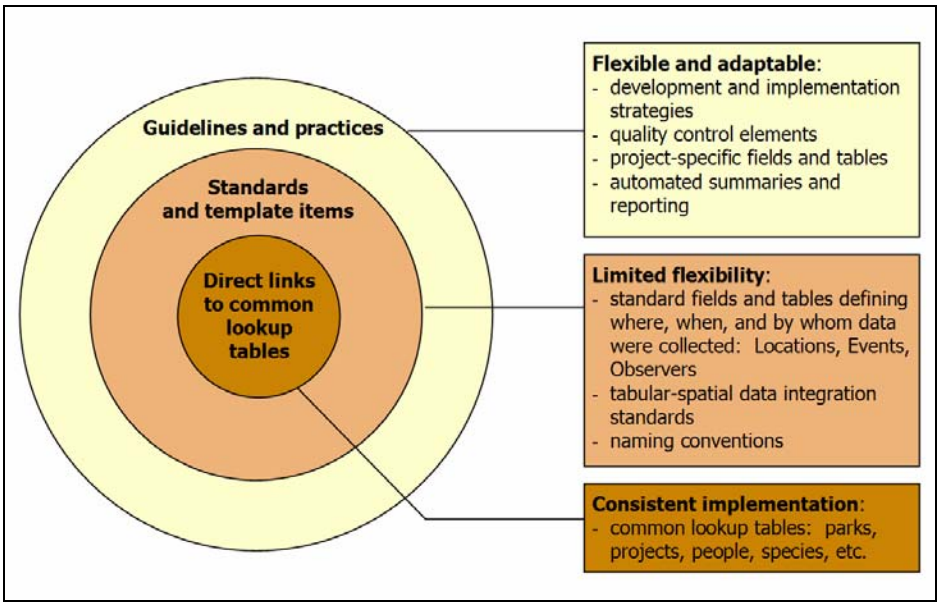

Figure 3.4. Different levels of data standards and their corresponding degree of implementation variability.

3.3.2 Project Tracking Application

To support program coordination and annual reporting, and to improve accountability for the products of our natural resource inventory and monitoring efforts, our Network will develop and implement a project tracking database. The primary functions of this application include:

- *Maintaining the list of projects* – a single list of natural resource data projects makes it much easier to quickly find project-related information (*e.g.*, status, funding sources and amounts, objectives, contact information) and summarize that information for administrative reports.
- *Tracking products* – for each project a comprehensive list is maintained of what products are expected and when they are due. Once they are delivered and posted or archived, this function

shifts to a navigational finding aid for available products. Expected products are first specified at project initiation and information is updated at various project milestones (*e.g.*, contracting, product delivery, archival).

- *Managing project codes* – these are intelligent alphanumeric codes used to tie together digital information in various, minimally connected systems (*e.g.*, RPRS, PMIS), along with analog materials that cannot otherwise be linked to an integrated information system. These codes are also used to link to data in databases and GIS themes, especially where information from multiple sources is stored together.

This application will be hosted by the network data server. Although primarily maintained by the Network data manager, the database will be available to project leaders, GIS staff, the network coordinator, and other network administrators. Each of these staff will be able to make certain changes to update information about project status, product details, etc. Certain database views will be created to help project leaders keep on schedule, and to facilitate quick reporting on project status, accomplishments, and delivered products.

This section will be updated once the application is implemented and additional details are available.

Credits

This chapter was adapted from material developed by John Boetsch (North Coast and Cascades Network), Lisa Nelson (WASO; section 3.2), and Patrick Flaherty (Appalachian Highlands Network).

Chapter 4. Data Management Process and Work Flow

This chapter considers the general work flow characteristics of projects that produce natural resource data, and then gives an overview of how natural resource data are generated, processed, finalized, and made available. Data management activities that relate to the various stages of a project are highlighted. By describing the progressive stages of a project and the life cycle of the resulting data, we can more easily communicate the overall objectives and specific steps of the data management process. In addition, this awareness helps us to manage the staffing resources needed to produce, maintain, and deliver quality data and information. More details about data acquisition, quality assurance, documentation, dissemination, and maintenance can be found in later chapters of this plan.

4.1 Project Work Flow

From the perspective of managing work flow, there are two main types of projects:

- *Short-term projects*, which may include individual park research projects, inventories, or pilot studies conducted in preparation for long-term monitoring.
- *Long-term projects*, which will mainly be the implemented monitoring projects central to the SODN I&M Program, but which may also include multi-year research projects and monitoring performed by other agencies and cooperators. Long-term projects will often require a higher level of documentation, peer review, and program support.

From a data management standpoint, a primary difference between short-term and long-term projects is an increased need to adhere to standards for long-term projects to ensure internal compatibility over time. This does not, however, negate the need to follow standards for short-term projects whenever possible and practical. Both short- and long-term projects share many work flow characteristics, and both generate data products that must be managed and made available.

Projects can be divided into five primary stages: planning and approval, design and testing, implementation, product integration, and evaluation and closure (Figure 4.1). Each stage is characterized by a set of activities carried out by staff involved in the project. Primary responsibility for these activities rests with different individuals according to the different phases of a project. Additional discussion of the different roles and responsibilities of park and network staff can be found in Chapter 2 of this plan.

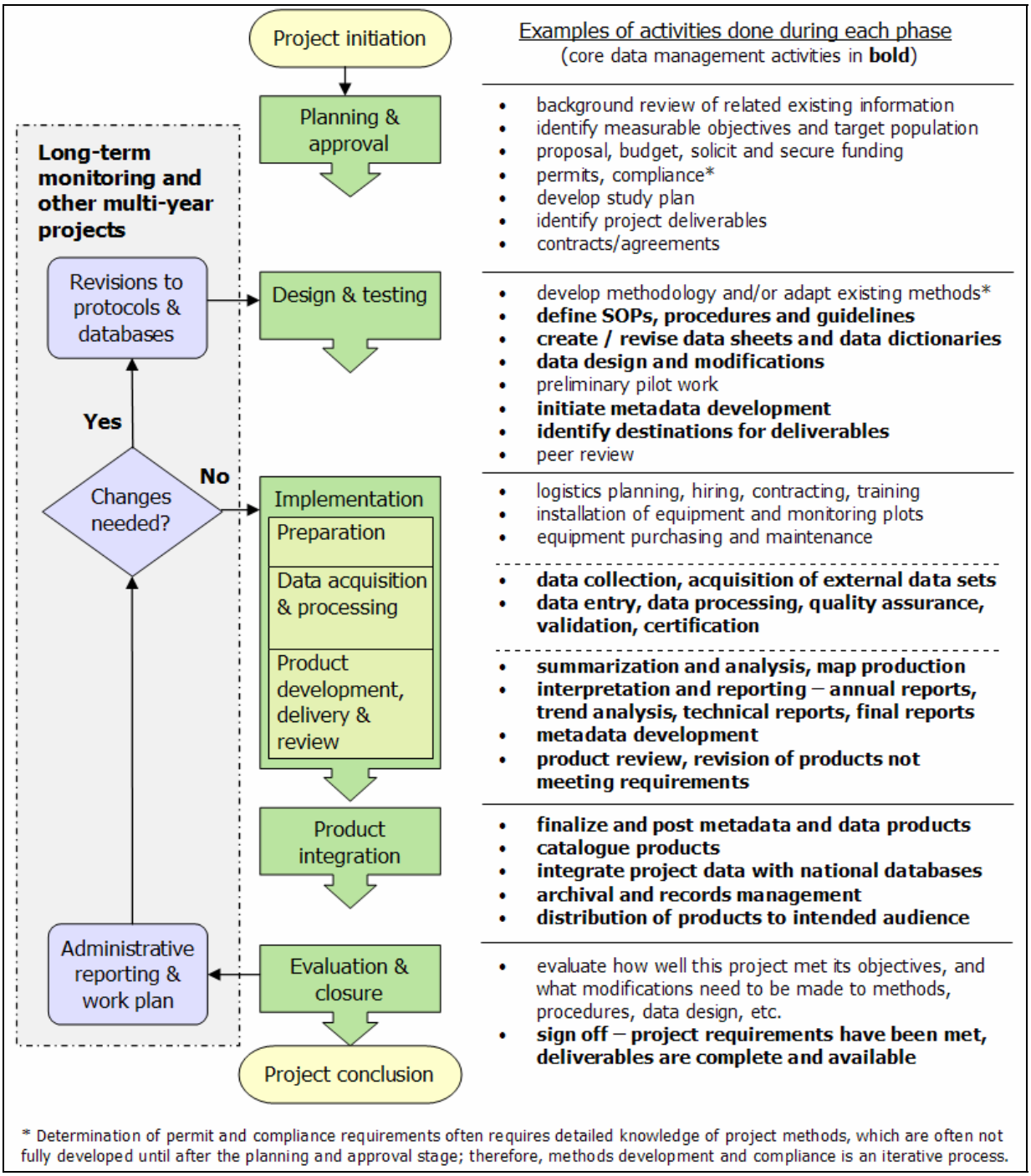

Figure 4.1. Conceptual model of project work flow.

Planning and Approval

During this initial phase, many of the preliminary decisions regarding project scope and objectives are made, and funding sources, permits, and compliance are all addressed. Primary responsibility rests with project leaders and program administrators. Although this phase lacks specific data management activities, it is important that data managers remain informed of projects in this phase. This is especially true as timelines for interim and final products are finalized. A 'project notification' form will be completed when a project is initiated and given to the data manager. All contracts, agreements, and

permits should include standard language that describes the formats, specifications, and timelines for products resulting from the project.

Design and Testing

During this phase, all of the details are worked out regarding how data will be acquired, processed, analyzed, reported, and made available to others. The project leader is responsible for developing and testing project methodology or for modifying existing methods to meet project objectives. It is critical that the project leader and the data manager work together throughout this phase. The dialog between these two will help to build and reinforce good data management throughout the project, especially during the crucial stages of data acquisition, processing, and retrieval. By beginning collaborative development as soon after project approval as possible, data integrity and quality can most easily be assured. An important part of this collaboration is the development of the data design and data dictionary, where the specifics of database implementation and parameters that will be collected are defined in detail. Devoting adequate attention to this aspect of a project is possibly the single most important part of assuring the quality, integrity, and usability of the resulting data. Once the project methods, data design, and data dictionary have been developed and documented, a database can be constructed to meet project requirements.

Implementation

During the implementation phase, data are acquired, processed, error-checked, and documented. This is also when products such as reports, maps, GIS themes, and others are developed and delivered. The project leader oversees all aspects of implementation – from logistics planning, contracting, training, and equipment procurement to data acquisition, report preparation, and final delivery. Throughout this phase, data management personnel function primarily as facilitators by providing training and support for database applications, GIS, GPS, and other data processing applications; facilitation of data summarization, validation, and analysis; and assistance with the technical aspects of documentation and product development. The specific roles of data management staff during this phase will depend primarily on the technical capabilities of the project staff. As much as is possible, these roles should be worked out in advance of implementation.

Toward the end of this phase, project staff members work to develop and finalize the products that were identified in the project planning documents (*i.e.*, protocol, study plan, contract, agreement, or permit). In general, all raw and derived data products, metadata, reports, and other documentation should be delivered to the project leader (data steward) assigned to the project for review. After the project leader's review, the products are delivered to the data manager. Administrative records should be delivered to appropriate park and network staff as specified. All project products should be developed and delivered according to product specifications, which should be stipulated in all protocols, contracts, agreements, and permits. Products that do not meet program requirements will be returned for revision.

Product Integration

During this phase, data and other products are integrated into national and network databases, metadata records are finalized and posted in clearinghouses, and products are distributed or otherwise made available to their intended audience. Another aspect of integration is merging data from a working database to a master database maintained on the network server. This occurs only after the annual working data set has been certified for quality by the project leader. Certain projects may also have additional integration needs, such as when working jointly with other agencies for a common database.

Product integration includes creating records for reports and other project documents in NatureBib, posting imaged documents to the appropriate repository, posting metadata records that have been completed and submitted by project leaders, and updating NPSpecies to reflect any new species

occurrence information derived from the project. This will allow the information from the project to be searchable and available to others via service-wide search engines.

Evaluation and Closure

Upon project closure, records are updated to reflect the status of the project and its associated deliverables in a network project tracking application. For long-term monitoring and other cyclic projects, this phase occurs at the end of each field season and leads to an annual review of the project. For non-cyclic projects, this phase represents the completion of the project. After products are cataloged and made available, program administrators, project leaders, and data managers should work together to assess how well the project met its objectives and to determine what might be done to improve various aspects of the methodology, implementation, and formats of the resulting information. For monitoring protocols, careful documentation of all changes is required. Changes to methods, SOPs, and other procedures are maintained in a tracking table associated with each document. Major revisions may require additional peer review.

4.2 Data Life Cycle

During various phases of a project, the data take on different forms and are maintained in different places as they are acquired, processed, documented, and archived. This data life cycle is characterized by a sequence of events that we can model to facilitate communication. These events involve interactions with the following objects:

- *Raw data* – analog data recorded by hand on hard-copy forms and digital files from handheld computers, GPS receivers, automated data loggers, etc.
- *Working database* – a project-specific database for entering and processing data for the current season (or other logical period of time). This might be the only database for short-term projects where there is no need to distinguish working data for the current season from the full set of validated project data.
- *Certified data and metadata* – completed data and documentation for short-term projects, or one season of completed data for long-term monitoring projects. Certification is a confirmation by the project leader that the data have passed all quality assurance requirements and are complete and ready for distribution. Metadata records include the detailed information about project data needed for their proper use and interpretation (see Chapter 7).
- *Master database* – project-specific database for storing the full project data set, used for viewing, summarizing, and analysis; only used to store data that have passed all quality assurance steps.
- *Reports and data products* – information that is derived from certified project data.
- *Edit log* – a means of tracking changes to certified data.
- *National databases and repositories* – applications and repositories maintained at the national level, primarily for the purpose of integration among NPS units and for sharing information with cooperators and the public.
- *Local archives and digital library* – local storage of copies of data, metadata, and other products generated by projects. Archives are for hard-copy items and off-line storage media, whereas the digital library is maintained live on a server.

Although the data life cycle may vary depending on specific project needs and objectives, the typical life cycle for Network projects proceeds as follows (Figure 4.2):

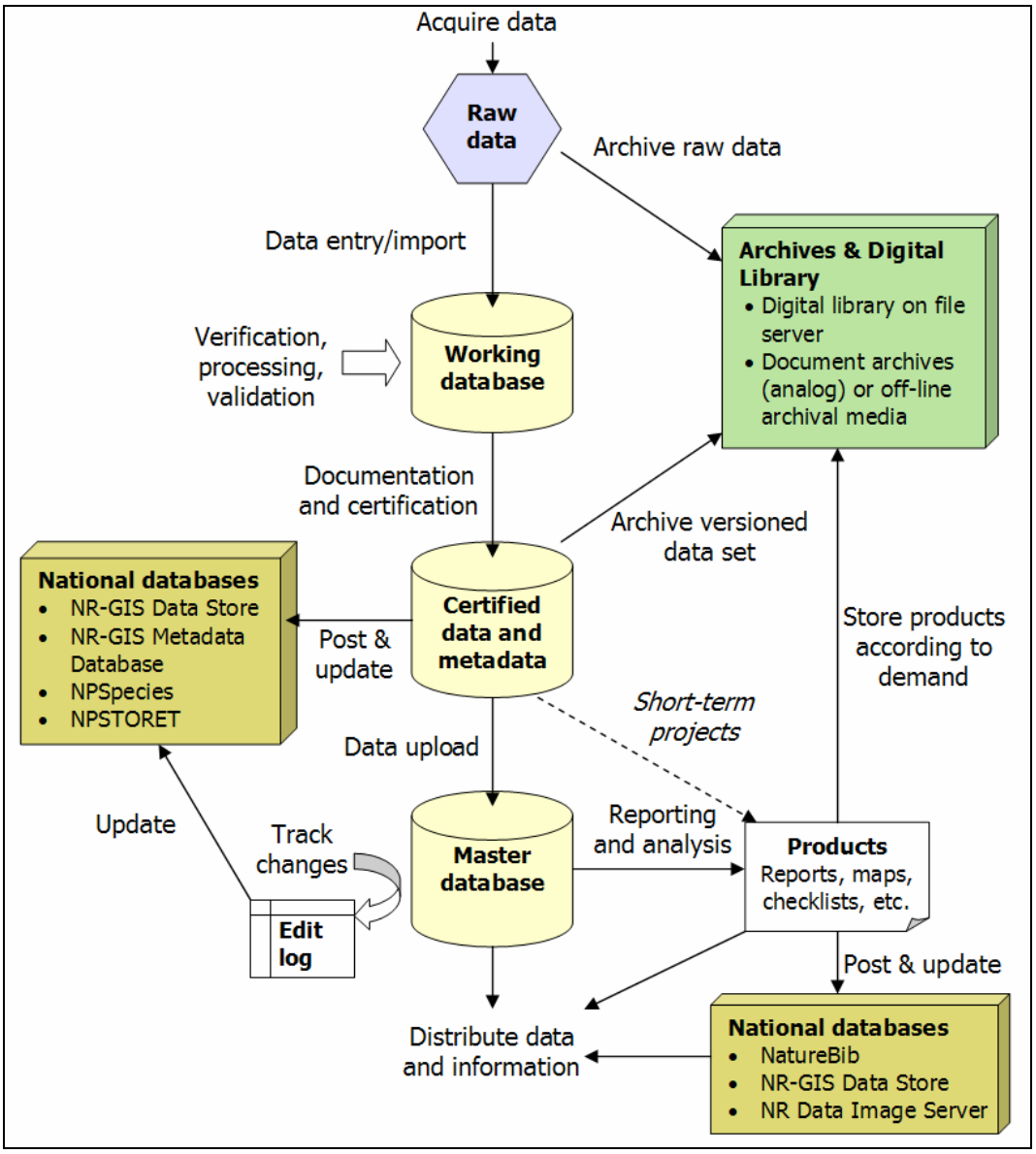

Figure 4.2. Diagram of the typical project data life cycle.

1) *Acquire data* – for data recorded by hand in the field, data forms should be reviewed regularly (at least daily) for completeness, legibility, and validity in order to capture errors as close to their origin as possible.
2) *Archive raw data* – copies of all raw data files are archived intact. Digital files are copied to the raw data folder for the project; hard copy forms are either scanned and placed in the active projects folder or are copied and placed in the archives. Archival or scanning of hard copy data forms may occur at the end of a season as a means of retaining all marks and edits made during the verification and validation steps.
3) *Data entry/import* – analog data are entered manually; digital data files are uploaded to the working database.
4) *Verification, processing, and validation* – verify accurate transcription of raw data; process data to correct data entry errors and remove missing values and other data flaws; validate data using database queries and other methods to capture missing data, out-of-range values, and logic errors.

5) *Documentation and certification* – develop or update project metadata and certify the data set. Certification is a confirmation by the project leader that the data have passed all quality assurance requirements and are complete and documented. It also means that data and metadata are ready to be posted and delivered.
6) *Archive versioned data set* – copies of the certified data and metadata are placed in the project archive folder. This can be accomplished by storing a compressed copy of the working database or by exporting data to a more software-independent format (*e.g.*, ASCII text; see Chapter 10).
7) *Post data and update national databases* – to make data available to others, certified data and metadata are posted to national repositories such as the NR-GIS Data Store. In addition, national databases such as NPSpecies, NPSTORET, and the NR-GIS Metadata Database are updated. **Note:** Data and data products may not be posted on public sites if they contain protected information about the nature or location of rare, commercially valuable, threatened or endangered species, or other natural resources of management concern (see Chapter 9).
8) *Upload data* – certified data are uploaded from the working database to the master project database. This step might be skipped for short-term projects where there is no need to distinguish working data for the current season from the full set of certified project data.
9) *Reporting and analysis* – certified data are used to generate data products, analyses, and reports, including semi-automated annual summary reports for monitoring projects. Depending on project needs, data might be exported for analysis or summarized within the database.
10) *Store products* – reports and other data products are stored according to format and likely demand – either in the digital library, on off-line media, or in the document archives.
11) *Post products and update national databases* – to make data available to others, reports and other products are posted to national repositories such as the NR-GIS Data Store or the NR Data Image Server. In addition, products are cataloged in NatureBib. Data products may not be posted on public sites if they contain protected information about the nature or location of rare, commercially valuable, threatened or endangered species, or other natural resources of management concern (see Chapter 9).
12) *Distribute data and information* – data, metadata, reports, and other products can be shared and distributed in a variety of ways – via the web-based national databases and repositories, by FTP or mailing in response to specific requests, or by providing direct access to project records to cooperators. In all cases, distribution will follow legal requirements under the Freedom of Information Act, and limitations will be established to protect information about sensitive resources (see Chapter 9).
13) *Track changes* – all subsequent changes to certified data are documented in an edit log, which accompanies project data and metadata upon distribution. Significant edits will trigger reposting of the data and products to national databases and repositories.

This sequence of events occurs in an iterative fashion for long-term monitoring projects, whereas the sequence is followed only once for short-term projects. For projects spanning multiple years, decision points include whether or not a separate working database is desirable and the extent to which product development and delivery is repeated year after year.

4.3 Integrating and Sharing Data Products

Once project data and derived products have been finalized, they need to be secured in long-term storage and made available to others. We use a range of information systems such as product repositories, clearinghouses, and web applications to accomplish this. Each of these systems has a different purpose and function, as shown in Figure 4.3.

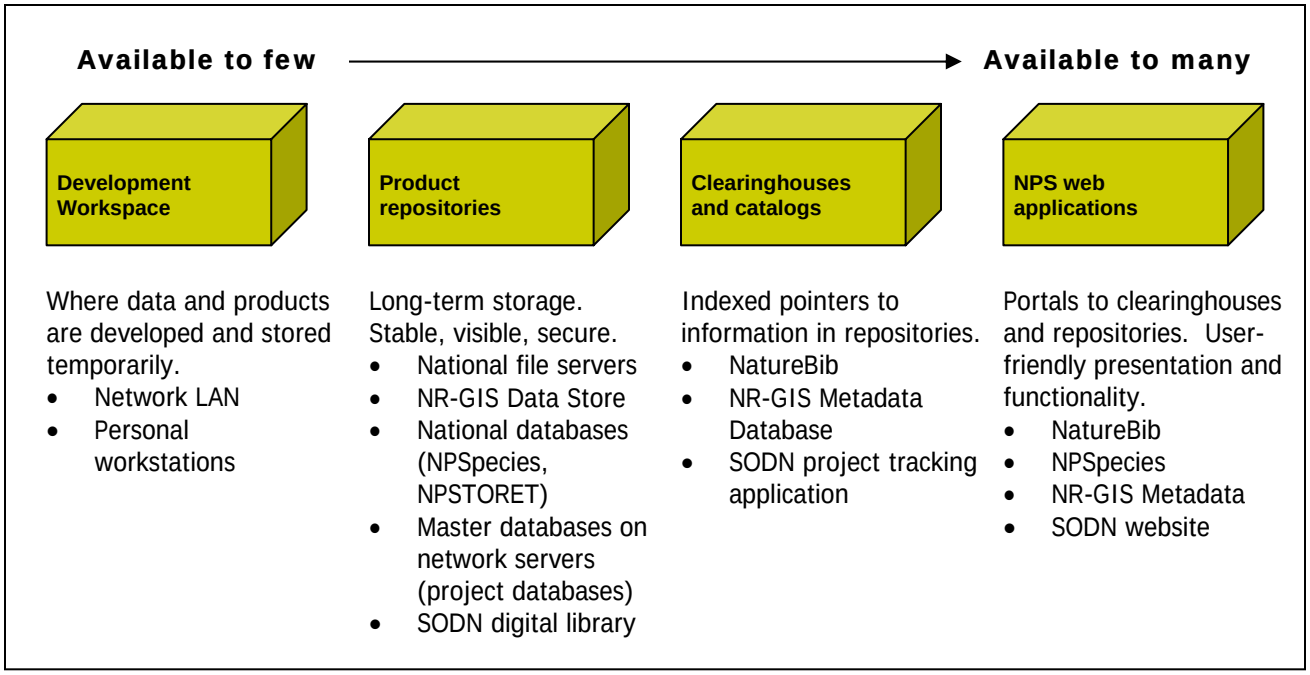

Figure 4.3. Storing and disseminating project information.

The specific repositories for most Network products are indicated in Table 4.1.

Table 4.1. Repositories for Network products.

Item	Repository
Reports	SODN project archive folder; posted to NR Data Image Server, linked and accessed through the catalog record in NatureBib; SODN library and park collection (hard copy)
Digital data sets (non-sensitive)	NR-GIS Data Store Biodiversity Data Store
Digital data, metadata, and other products Raw and finalized data Metadata, protocols, SOPs Digital photographs, derived products	SODN data servers and digital library; other cooperators for selected monitoring projects: Arizona Game and Fish Department, US Environmental Protection Agency, US Forest Service, etc.
Project materials Voucher specimens, raw data forms	Network and park archives and collections, or another specified collection (*e.g.*, Western Archeological and Conservation Center)
Administrative records	SODN offices and/or park offices, park archives, National Archives

4.3.1 Data Distribution

The process of product distribution involves several steps (see example for a GIS data set in Figure 4.4). As products are finalized, they can be sent to the appropriate person for integration, posting, and distribution. In most cases it will be either the data manager or GIS specialist who reviews the product

for conformance with format standards, then stores the product in the appropriate repository. Note that it is expected that all products will have already been reviewed for completeness and accuracy prior to delivery. After storing the products, their existence is documented by posting metadata and by updating records in the Network's project tracking application. At this point, data discovery is possible as metadata are then indexed by the clearinghouse function of the NR-GIS Metadata Database. These metadata records provide pointers to data and data products. Distribution follows data discovery when potential users find and either request or download the data sets from their repositories.

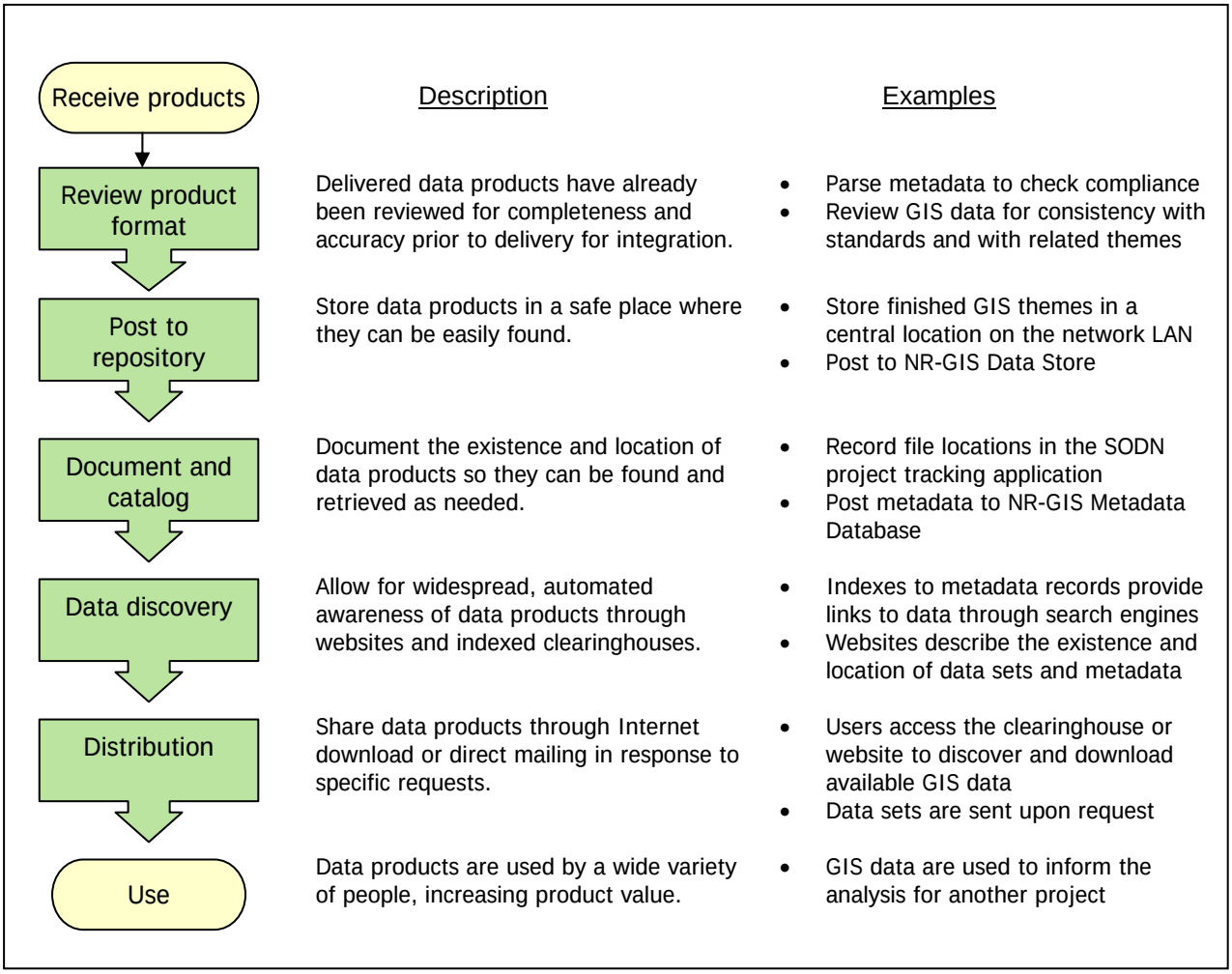

Figure 4.4. Steps involved in product distribution.

4.3.2 Integration with National Databases

In addition to storing and distributing data products, product integration also involves updates to national databases such as NPSpecies and NPSTORET. Both of these databases have local desktop versions that can be updated with data collected during the course of a project. Desktop databases are then uploaded and synchronized with the national databases on a regular basis.

To update NPSpecies, data on the distribution and occurrence of species in Network parks will be compiled and added to the database upon delivery of data and data products. Updates will be performed in the master online version whenever possible. If the desktop version is used, synchronization with the

master version of NPSpecies will occur at least twice annually, or more frequently depending on the timing and amount of updates.

For NPSTORET, any project collecting water quality data will be flagged in the project tracking application so that water quality data can be either extracted and uploaded or directly entered into NPSTORET. All water quality data collected by the Network will be managed according to guidelines from the NPS Water Resources Division. We will implement and maintain a desktop copy of NPSTORET and transfer its contents at least annually to NPS Water Resource Division for upload to the STORET database (Figure 4.5).

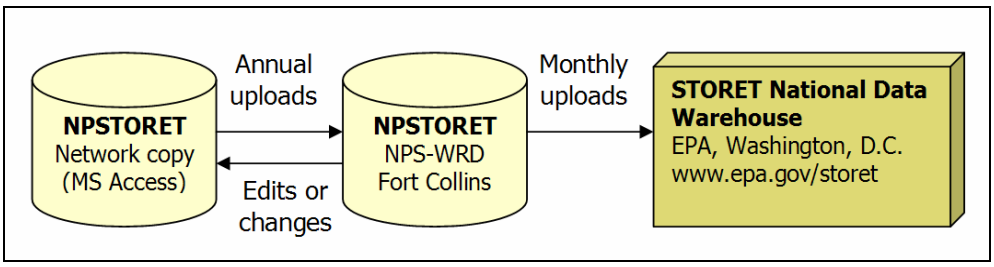

Figure 4.5. Data flow diagram for water quality data.

Credits

This chapter was adapted from concepts and material developed by John Boetsch (Northeast Coast and Cascades Network) in collaboration with Dorothy Mortenson (Southwest Alaska Network), Velma Potash (Cape Cod National Seashore), Sara Stevens (Northeast Coastal and Barrier Network), and Doug Wilder (Central Alaska Network).

Chapter 5. Data Acquisition and Processing

The NPS I&M Program, in support of the Natural Resource Challenge, is responsible for acquiring the information park managers need to manage and maintain their parks' natural resources. To successfully accomplish this task, we collect information from multiple sources and process it appropriately so that it meets established data standards. This chapter of the Data Management Plan describes the steps involved in acquiring data as well as selected stages of data processing. These steps are integral to providing high-quality data to strengthen the scientific foundation of the I&M Program and to enable park staff to effectively manage their natural resources.

Data handled by the SODN I&M Program fall into three general classifications that serve to differentiate among data acquired from different sources:

- *Programmatic data* are produced from projects that are either initiated (funded) by or substantially involve the I&M Program.
- *Non-programmatic NPS data* are produced by the NPS but did not involve the I&M Program.
- *Non-programmatic external data* are produced by agencies or institutions other than the NPS.

The importance or value placed on a data set in any of these categories will be based on its quality, completeness, relevance, and potential usefulness, as well as the impact it has on the SODN I&M Program and parks. Resource specialists can help determine quality, relevance, and usefulness. The data manager assists in determining quality and completeness.

The following sections outline in more detail the manner in which the different data types described above are acquired and processed.

5.1 Programmatic Data

We acquire program data most commonly through natural resource inventories or vital signs monitoring projects, but they may also be acquired through collaborative efforts between the SODN I&M Program and other regional or local programs and institutions. Data for these projects are typically collected by network personnel, park staff, or cooperators/contractors.

Natural resource inventories are designed to identify the primary resources of each park, and as such they provide important baseline data for the management of park resources and the development of the long-term monitoring program.

Monitoring studies are designed to collect data on vital signs; *i.e.*, specific, measurable ecological parameters that provide insights into broader ecological processes in natural systems. Repeated measurements of vital signs are performed to identify trends in resource conditions that indicate the effects of management actions and facilitate adaptive management. A negative trend can be an early warning signal that indicates the long-term functioning of the system is somehow impaired. Early detection of potential problems gives park managers the opportunity to take steps to reverse a downward trend in park resources before more serious consequences occur. Conversely, a positive or 'level' trend may indicate that current management actions are 'on the right track'; *i.e.*, helping to restore or maintain ecosystem integrity. Working in conjunction with park resource managers and local scientists, network staff selected and prioritized the natural resource vital signs that were of the greatest concern to individual SODN parks and to the Network as a whole. This collaborative effort identified 25 vital signs that the Network plans to monitor (see Chapter 3 of the SODN Vital Signs Monitoring Plan [Mau-Crimmins *et al.* 2005] for a list of the selected vital signs).

5.1.1 Field Studies

Biological inventories and monitoring projects are the most common examples of field studies conducted by the SODN I&M Program. The data manager is responsible for ensuring that data collection, data entry, verification, validation, storage, and archiving for all field projects are consistent with SODN I&M Program standards. In addition to general standard operating procedures that define network-wide requirements, SOPs detailing procedures and/or methodologies specific to each monitoring protocol will be developed. The data manager will work closely with the project leader and network staff to develop the guidelines for project data management. These may range from detailing the proper usage of data entry forms or databases to outlining calibration procedures for automated data loggers. Refer to individual protocols for specific SOPs. Examples of protocol-specific SOPs are available at http://www1.nrintra.nps.gov/im/datamgmt/dmplanning.htm.

Methods of field data collection will vary by project, depending upon the parameters being measured. Paper field notebooks or data forms have been the primary methods for ecological data collection for many years. Although paper has advantages in terms of longevity and ease of use, it does not work well under some environmental conditions, and processing options are limited until the data are transferred to digital format. As an alternative to paper, several options for electronic data collection in the field are now available, including field computers, automated data loggers, and GPS receivers. Refer to protocol-specific SOPs for details on how the following methods may apply to individual projects.

Paper field data forms are the most common method of recording field data. Although inexpensive, more opportunities for errors exist during the data collection/data entry process. They also require neat, legible handwriting and very rigorous QA/QC.

Field computers increase data collection and data entry efficiency. Data can be downloaded directly from field computers to office desktops, thereby eliminating manual data entry. Fewer chances for error exist as QA/QC checks can be built into the database, but these devices may not be the optimal choice if copious amounts of notes or comments must be recorded in the field. In addition, these portable units are subject to environmental constraints such as heat, dust, and moisture. When handheld computers are used for data entry in the field, the data should be downloaded daily to avoid potential loss of information. Then, if a handheld unit fails during data collection, only the current day's data are lost. Batteries should be checked prior to a data collection trip, and they should be charged at the end of every field day. The use of a memory card that will store the data in case of damage to the unit or battery failure can prevent accidental loss of data. In addition, in case the unit becomes inoperable in the field, printed data sheets should always accompany field teams on data collection trips.

- *Handheld computers or Personal Digital Assistants (PDAs)* – the small size and relative low cost of these devices make them attractive options for collecting data in the field. Although they work well for small field projects, they are not powerful enough for large, data intensive field projects. PDAs can be customized to withstand a range of adverse environmental conditions fairly easily and inexpensively. Most run either Windows CE or Palm operating systems, which may require additional processing/programming to transfer/create the database structure in the field units.
- *Tablet PCs* – these units have the same properties as most laptops and provide the user with the convenience of a touch screen interface. They are bulkier, more expensive, and harder to customize for fieldwork than the PDAs but are more powerful. They work well for field projects that are very data intensive. Because these units run Windows XP (Tablet Edition), the project database can be directly transferred from desktop units to field units without additional programming steps.

Automated data loggers are mainly used to collect ambient information such as weather data or water quality information. Data loggers are an efficient method for recording continuous sensor data, but routine inspections are necessary, and environmental constraints, as well as power (*i.e.*, sufficient battery charge) and maintenance requirements, are potential pitfalls when using these instruments. Regular downloads should be required since physical memory is usually limited. Proper calibration is important, so field crews must receive proper training and refer to the equipment manuals and protocol SOPs outlining the procedures.

- *Permanently deployed devices* – these devices are often very expensive, and data must be retrieved and batteries changed on a regular basis. These intervals should be defined in the protocol.
- *Portable hand-held devices* – these units are deployed for sampling only during site visits. They are generally less expensive than permanently-deployed field units.

GPS receivers are often used during fieldwork in network parks to collect location information.

- *Handheld recreation-grade GPS units* are relatively inexpensive and are good for collecting general position information, but they are not recommended for obtaining high accuracy location information.
- *Mapping-grade GPS receivers* are good for collecting highly accurate (sub-meter) location information, but they are more expensive than recreation-grade units, and more training is required to use these units correctly. However, due to the small size of many Network parks, field crews should be encouraged to use these units, if available, to collect GPS data.

5.1.2 Data Processing

Data processing refers to the series of steps that transform the acquired data (often raw data) into information and finished products. These steps include data entry, verification, editing, validation, documentation, dissemination, and archiving. These topics are covered in detail in Chapters 6-10 of this plan; here we present processing steps to integrate the data we collect into our Program information base and the service-wide databases.

Programmatic data will be processed using the following applications:

Natural Resources Database Template

All field studies funded by the Network will have an associated Microsoft Access database developed collaboratively by the data manager and the project leader, and we have adopted the Natural Resources Database Template (NRDT) as the foundation for our databases. The NRDT is highly flexible and can be modified and customized for each project to meet the project objectives and the researcher's requirements. The database incorporates mechanisms such pick lists and validation rules for quality assurance purposes. The following are general guidelines for using the NRDT:

- Field crews/project staff enter all data into the specified database.
- All data must undergo QA/QC procedures. (See Chapter 6 of this document for more specifics relating to data verification and validation.)
- The project leader is responsible for the data set for the duration of a project. He/she is responsible for submitting the project data on a regular basis (at the end of the project for short-term projects; once a year for monitoring or other long-term projects) to the data manager. Refer to individual protocols for the procedures regarding data submissions.

- The data manager maintains the master copy of the database and updates it with verified/validated data received from the project leader.

Dataset Catalog

New data sets (spatial and non-spatial) will be entered into Dataset Catalog, which will be made available in a online version to integrate with the NR-GIS Metadata Database.

SODN Library/NatureBib

The Network will enter all new references resulting from network projects into the SODN library database and the NatureBib database, if applicable. References are first entered into the SODN library catalog; qualified document citations will be uploaded to NatureBib at least every six months.

NPSpecies/ANCS+/Biodiversity Data

NPSpecies is a National Park Service database developed by WASO to store, manage, and disseminate scientific information on the biodiversity of organisms in National Park Service units throughout the United States and its territories. The database is available in an on-line form (Oracle) or a desktop version (MS Access). For more specifics on NPSpecies please refer to the following web page: http://science.nature.nps.gov/im/apps/npspp/Discover.htm.

Four forms of evidence are used to document species existence in NPS units:

- *References* documenting species existence in parks should be entered into NatureBib and the listed species linked in NPSpecies to the publication or report.
- *Data sets* documenting species existence in parks should be entered into NatureBib and the listed species linked in NPSpecies to the data set.
- *Observations* made during biological inventories or monitoring studies are entered into the NPSpecies database.
- *Voucher specimens* collected during inventories or monitoring studies must be entered into NPSpecies and the ANCS+ cataloging system. The park in which the vouchers were collected technically owns the specimens and has the right to decide where the specimens will be stored.

5.1.3 Changes to Data Collection Procedures/Protocols

Changes to established data collection procedures are strongly discouraged unless there are acceptable, valid reasons for altering the methodologies. Ideally, all problems should be identified during the design and testing stages of the project and changes implemented prior to the collection of any field data. Protocols should attempt to identify any foreseeable issues that might occur as well as contingencies to address them. However, unforeseen problems that require procedure/protocol revision after data collection has begun will inevitably occur. Significant changes to protocols must be approved by the project leader and data manager. The project leader must evaluate the proposed changes and determine if additional peer review is required before accepting them.

Altering data collection procedures or protocols may also occur as a result of the comprehensive review that all monitoring protocols will undergo every five years. During the review, data are evaluated to determine if the current protocol has accomplished its goal. If we conclude that the protocol in its present form has not achieved the desired results, revisions should be recommended. Once again, all changes must be approved by the project leader and data manager.

5.2 Non-Programmatic NPS Data

A large percentage of data collected in Network parks are collected by park personnel involved in projects initiated at the individual park level or by other NPS regional or national programs. The data collected and products produced by such efforts provide a great deal of information about park natural resources and are therefore relevant to the mission of the I&M Program.

5.2.1 Park Data

All parks have information that may contribute to the development of the monitoring program. For example, some parks have funded their own inventories or tracked particular taxa (*e.g.*, threatened or endangered species) through time. Others may have historic photographs, maps, and voucher specimens that illustrate the state of natural resources at one point in time, which may indirectly help to track changes in the condition of resources over time.

Parks in the Network often use base funding or receive funding through NRPP (Natural Resources Protection and Preservation) programs to support park-level projects.

- *Park based biological inventories* – Network parks often conduct their own park-based inventory projects, the data from which can be used to supplement Network-level inventories conducted by the I&M Program.
- *Park-based monitoring projects* – parks also engage in park-level monitoring projects (such as vegetation and water quality), which produce information that is valuable when developing Network-level monitoring protocols.
- *Park and multi-park based projects* – other studies or projects conducted at the park or regional level that do not fall into one of the previous two categories (*e.g.*, restoration projects).

5.2.2 Regional and National Program Data

NPS regional and national programs support all of the parks within the Intermountain Region and are good sources for natural resources information.

- *Air* **–** national-scale programs collect data, maintain databases, assure data quality, and perform the trend analyses relevant to air quality issues in the Network. The SODN I&M Program will rely on the data analyses from these national scale monitoring networks to obtain trends for many of its air vital signs.
- *Fire Program* **–** data concerning the occurrence of fires within the Network are maintained at both individual park and regional levels. National databases such as Fire-Pro, SACS, and the soon-to-be implemented Fire Program Analysis (FPA) package (http://fpa.nifc.gov/) have been and will be used to maintain information regarding fire incidences and fire management resources. The NPS is also involved in efforts such as the Joint Fire Science Program (http://jfsp.nifc.gov/) that provides scientific information and support for fuel and fire management programs.
- *GIS* **–** the SODN I&M Program is supported by regional and national GIS specialists to help ensure that GIS data are available and accurate. Many of these data are available through the NR-GIS Metadata and Data Store and the Spatial Data Clearinghouse.
- *Water* – water resources monitoring efforts are taking place at network, regional, and national scales. Almost all of the field data collection by the regional water resources program is done in support of the water resources vital signs monitoring projects. The program also synthesizes, analyzes, and interprets water resources data collected by parks. The Network office houses a

regional hydrologist who provides expertise to network and park staff in Arizona and New Mexico, as well as selected parks in Colorado and Utah.
- *Wildlife management* – the regional wildlife biologist provides management support to the parks in the Sonoran Desert Network. The Wildlife Management Program coordinates long-term monitoring and assessments of wildlife populations.

5.2.3 Data Processing

Existing NPS data often do not require a great deal of processing because the SODN I&M Program shares many file standards with parks and regional programs. Basic processing steps include:

- Enter all new park biodiversity data into NPSpecies (especially important for park-based biological inventories), and enter all associated references into NatureBib.
- Enter all park-based or regional reports and publications related to natural resources in parks into the SODN Library database. Hard copies should be stored in the appropriate file cabinets and electronic copies archived in the proper directory on the network data server. At least every six months, qualified document citations will be uploaded to the online NatureBib database.
- Ensure that all GIS data include projection information files and are accompanied by FGDC-compliant metadata.
- Enter all data sets into Dataset Catalog for tracking purposes.

If existing data are in digital format, the data set will be converted to current standard formats compatible with current NPS software standards as time and resources permit. Analog data may be maintained as such or they may be imported into the NRDT. Selected analog publications and reports may be scanned to .pdf format and added to the digital library. Priority for conversion of legacy/existing data sets will be based on their relevance to the SODN I&M Program and parks and projected frequency of use.

It is important that information is maintained in a manner that promotes data sharing. Accordingly, the Network data manager will:

- Work closely with park and regional personnel to ensure that high-quality data are available.
- Provide training to park staff on the use of NPSpecies, NatureBib, Dataset Catalog, the NRDT, and GIS applications.
- Work with park staff to ensure relevant information collected/maintained by the parks is entered into NPSpecies or NatureBib.
- Provide assistance with the processing of voucher specimens collected in network parks.
- Assist park staff in developing databases based on the NRDT that meet the needs of park resource managers.

5.3 Non-Programmatic External Data

External (*i.e.*, non-NPS) agencies and institutions often possess relevant data regarding natural resources in the Sonoran Desert region. Such information need not be directly related to network parks but may instead pertain to topics such as sampling methodologies, natural resource management on lands adjacent to NPS units, or the general ecology of the region. Nevertheless, this information could assist network personnel with the development of the monitoring program. For example, the Network does not collect data for monitoring air quality vital signs; we rely on external sources for this information. These data will be obtained from national networks such as the Clean Air Status and Trends Network (CASTNet) for ozone monitoring, the National Atmospheric Deposition Program (NADP), and Interagency Monitoring of Protected Visual Environments (IMPROVE). Network staff will periodically download and archive

the data and accompanying metadata according to the SOPs provided by the NPS Air Resources Division on their monitoring website (http://www2.nrintra.nps.gov/ard/monitoring/index.htm).

The following sources may have information pertinent to the SODN I&M Program:

- Federal, State, and/or local (county or city) government agencies (*e.g.*, Bureau of Land Management, U.S. Forest Service, Pima County)
- Academia (*e.g.*, University of Arizona, Arizona State University)
- Private organizations/non-profit groups (*e.g.*, Sonoran Institute, The Nature Conservancy)

The agencies or organizations that compile these data have the expertise to apply proper quality control procedures and the capability to function as a repository and clearinghouse for the validated data. When the data are not kept in-house, data may be acquired via downloads from online databases or requests for data on CD, DVD, or other media.

5.3.1 Data Processing

Unlike data from NPS sources, much of the data collected from external sources must undergo some degree of processing to meet the standards of the SODN I&M Program; however some of the basic processing steps are very similar.

- All GIS data obtained from other entities will be stored in the proper format and include accurate spatial reference information and FGDC-compliant metadata. This is especially applicable to data collected in support of the vital signs relating to land use/land cover change.
- All biodiversity data received from other entities, such as Breeding Bird Survey data sets, should be entered into NPSpecies. In addition, if the data were taken from a report or published document, the reference must be entered into NatureBib.
- All reference materials obtained will be entered into the SODN library database, and qualified document citations will be uploaded to NatureBib at least every six months.
- All data sets will be entered into and tracked using Dataset Catalog.

Certain data sets will require more than the basic processing steps described above. The level of data processing required for external data sets such as those used in the vital signs monitoring program depends on the desired output. Additional processing steps may be required if the network intends to conduct more intensive analyses than those provided from these sources. For example, if an in-depth analysis at a specific location rather an overall regional trends analysis is desired, the data may require further processing. Another example is a data set that will be used in interdisciplinary analyses. This data set may have to be imported into the NRDT format so it can be integrated with network data. In such cases, the specific protocols should outline the necessary data processing requirements.

Remote sensing data sets (*e.g.*, satellite imagery or aerial photography) may require geospatial or spectral processing, depending upon the formats in which they are received. Ideally, all spatial data sets will be received in a geo-referenced format and may require only geographic transformations to meet Network standards. Varying degrees of spatial and spectral processing may be necessary to adequately answer the proposed questions. Again, protocol-specific SOPs should provide the necessary data processing requirements for data from external sources.

5.4 Data Discovery/Data Mining

Data discovery or data mining is the process of searching for existing natural resource related data and information from NPS and external sources that may be useful to the SODN I&M Program. Although

historically the NPS has not always built upon existing information when implementing new studies, considerable time and attention in the development of vital signs monitoring has been focused toward creating a monitoring program that integrates existing information and builds on an information base. Data mining contributes to the information foundation upon which the vital signs monitoring program is built.

From the perspective of the Network data management system, data mining is defined as the collection, cataloging, consolidation, and organization of existing natural resource information pertinent to the network parks. This definition of data mining differs from the standard IT notion of data mining, which focuses on extracting information from large digital databases.

Much of the information collected during the data mining process is likely to be *legacy data*, or data collected prior to the inception of the SODN I&M Program. If legacy data are collected in a digital format, the information should be converted to current file formats compatible with current software standards as time and resources permit. Hard copy references and reports may be scanned and saved as .pdf files in order to create a digital library.

Even though data discovery efforts most often occur at the initiation of new projects or during the development of new protocols and are an integral part of project development, data mining should not be limited solely to project development needs. It should be an ongoing process entailing periodic data searches and visits to network parks to ensure that the SODN I&M Program maintains as much relevant material pertaining to park resources as possible. Encouraging data sharing with network parks will facilitate this process and may reduce the need for regular searches of park records. The SODN I&M Program plans to conduct general data discovery efforts as an ongoing effort.

A large percentage of data discovery occurs at the onset of new projects or during the development of new protocols. Data mining involves reviewing many different sources for varying types of information. Many of the following data sources are accessible via the Internet, but some can only be accessed through visits to local libraries, academic institutions, museums, or nearby parks.

Bibliographic/Literature

- National NPS databases (*e.g.*, NatureBib)
- Online literature databases (*e.g.*, First Search, Biosis)
- Local document library (*e.g.*, SODN library)
- Library catalogs (academic or research institutions)
- Park archives, libraries, and files

Geographic Data

- Regional GIS specialists
- Park GIS specialists
- Federal and State geographic data clearinghouses

Biological/Natural Resources Data

- National NPS databases (*e.g.*, NPSpecies)
- Voucher collections (museums, parks, universities)
- Network parks
- Federal, state, and local government land management agencies

All ‘mined’ data sets should be evaluated for their potential contribution to the development of the Vital Signs Monitoring Program. Suggested criteria for evaluating the usefulness of data sets include:

- Sample design and methods are clearly defined
- Data are verifiable
- Data are temporally registered
- Data are spatially registered
- Data fields are defined, including units of measure
- Species nomenclature is defined
- Data are relevant to ongoing or planned future monitoring

All information collected during the data discovery process is maintained either electronically or in hard copy format, depending on how it was collected. Any geographic data sets collected during this process should be accompanied by FGDC-compliant metadata, and all data sets found – geographic or otherwise – should be entered into Dataset Catalog.

Dataset Catalog – new data sets (spatial or non-spatial) should be entered into Dataset Catalog.
SODN Library – the desktop version of NatureBib has been adopted to function as the internal literature library catalog for the Network. All natural resource reference information collected during the data discovery process is entered into this library catalog and will be uploaded to the online NatureBib database at least every six months.
NPSpecies/NatureBib – information relating to the biodiversity of Network parks is entered into NPSpecies and linked to the associated reference (assuming the data came from a report or publication) or data set in NatureBib.
Library Reference Cabinets – hard copy materials are stored in one or more locations depending upon how they are related to Network parks and the SODN I&M Program. Original data sheets, final reports, and contracts are stored in [*repository to be determined*]. General reference materials, including those linked to NPSpecies records in NatureBib are stored in file cabinets in the SODN library.

Credits

This chapter was adapted from material developed by Geoffrey Sanders (National Capital Region Network).

Chapter 6. Data Quality Assurance / Quality Control

We consider the ecological data and information collected during inventories and monitoring studies as valuable resources that must be preserved over the long-term. However, this perception is justified only if we have confidence in our data. Our analyses to detect trends or patterns in ecosystem processes require data of documented quality that minimize error and bias. Data of inconsistent or poor quality can result in loss of sensitivity and lead to incorrect interpretations and conclusions. The potential for problems with data quality increases dramatically with the size and complexity of the data set (Chapal & Edwards 1994).

Palmer (2003) defines *Quality Assurance* (QA) as "an integrated system of management activities involving planning, implementation, documentation, assessment, reporting, and quality improvement to ensure that a process, item, or service is of the type and quality needed and expected by the consumer." He defines *Quality Control* (QC) as "a system of technical activities to measure the attributes and performance of a process, item, or service relative to defined standards." QA procedures maintain quality throughout all stages of data development; QC procedures monitor or evaluate the resulting data products.

This chapter presents the procedures the SODN I&M data management system will follow to ensure that our projects produce and maintain data of the highest possible quality. We will establish and document protocols for the identification and reduction of error at all stages in the data life cycle. Although a data set containing no errors would be ideal, the cost of attaining 95%-100% accuracy may outweigh the benefit. Therefore, we consider at least two factors when setting data quality expectations:

- Frequency of incorrect data fields or records
- Significance of error within a data field

We are more likely to detect an error when we work with clearly documented data sets and understand what a 'significant' error is within *that* data set. The significance of an error can vary among data sets and depends on where it occurs. For example, a two-digit number off by one decimal place is a significant error. A six-digit number, with the sixth digit off by one decimal place, is not a significant error. However, one incorrect digit in a six-digit species number could indicate a different species. That is a significant error.

6.1 National Park Service Quality Mandate

Not long ago, maintaining data meant filling filing cabinets full of notebooks and paper. Now we are more likely to use computer hardware and software – technology that changes rapidly and sometimes unpredictably. If we expect our current data to be useful to future users, the data must survive changes in technology. We can promote data longevity through high-quality documentation and maintenance during all phases of data management. Well-documented data sets are especially important when sharing data.

NPS Director's Order #11B: "Ensuring Quality of Information Disseminated by the National Park Service," issued in 2002, promotes information and data quality. It defines *'quality'* as incorporating three key components – *objectivity*, *utility*, and *integrity*.

- *Objectivity* consists of: 1) *presentation,* which focuses on whether disseminated information is being presented in an accurate, clear, complete, and unbiased manner within a proper context, and 2) *substance,* which focuses on ensuring accurate, usable, and reliable information.
- *Utility* refers to the usefulness of the information to its intended users, from the perspectives of the Network, resource managers, other scientists and users, and the general public.

- *Integrity* refers to the security of information, *i.e.*, protection from unauthorized access or revision to ensure that the information is not compromised through corruption or falsification.

DO #11B also specifies that information (*e.g.*, brochures, research and statistical reports, policy and regulatory information, and general reference information) must be based on reliable data sources, which are accurate, timely, and representative of the most current information available. These standards apply not only to NPS-generated information, but also to information provided by other parties to the NPS if the NPS disseminates or relies upon this information.

Not only are high-quality data and information mandated by directives and orders, they are vital to the credibility and success of the I&M Program. According to Abby Miller (2001) of the Natural Resource Stewardship and Science Division,

> ". . . data need to meet national-level quality standards and need to be accessible to be used for wise and defensible decision-making at all levels. Data need to be able to be shared and aggregated with data from other parks and from adjacent lands to support landscape-level and national planning and decision-making."

6.2 Quality Assurance and Quality Control Mechanisms

QA/QC mechanisms are designed to prevent data contamination, which occurs when a process or event other than the one of interest affects the value of a variable. Contamination introduces two fundamental types of errors into a data set:

- *Errors of commission* include those caused by data entry and transcription errors or malfunctioning equipment. They are common, fairly easy to identify, and can be effectively reduced upfront with appropriate QA mechanisms built into the data acquisition process, as well as QC procedures applied after the data have been acquired.
- *Errors of omission* often include insufficient documentation of legitimate data values, which may affect the interpretation of those values. These errors may be harder to detect and correct, but many of these errors should be revealed by rigorous QC procedures.

QA/QC procedures applied to ecological data include four procedural areas (or activities), ranging from simple to sophisticated, inexpensive to costly:

1) Defining and enforcing standards for electronic formats, locally defined codes, measurement units, and metadata
2) Checking for unusual or unreasonable patterns in data
3) Checking for comparability of values between data sets
4) Assessing overall data quality

Much QA/QC work involves the first activity (defining and enforcing…), which begins with data design and continues through acquisition, entry, metadata development, and archiving. The progression from raw data to verified data to validated data implies increasing confidence in the quality of the data through time (Figure 6.1).

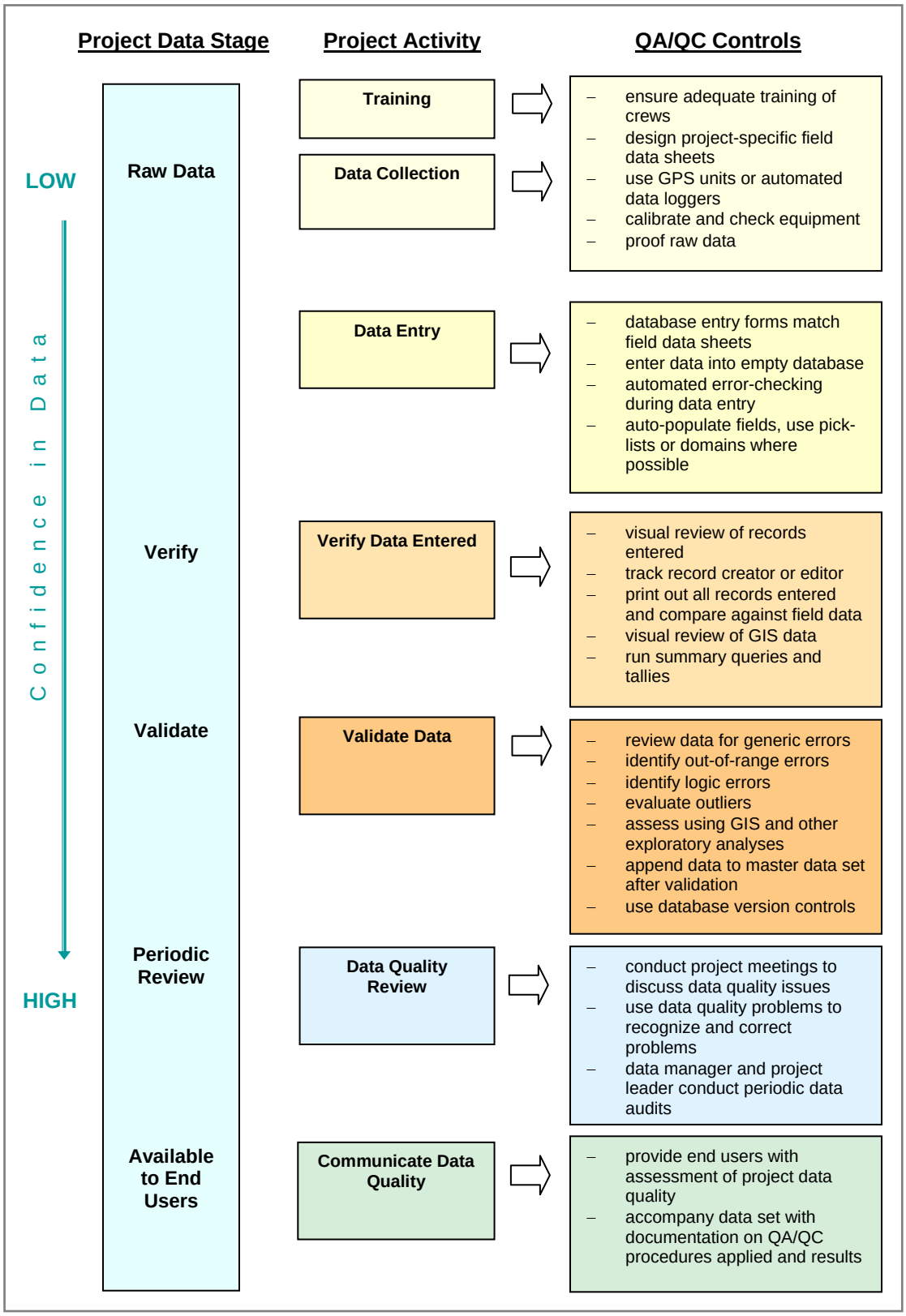

Figure 6.1. QA/QC controls applied at progressive stages of a project.

6.3 Roles and Responsibilities

Quality assurance methods should be in place at the inception of any project and continue through all project stages to final archiving of the data set. It is critical that each person involved with the data works to ensure data quality. Everyone plays a part in producing and maintaining high-quality data, and everyone assigned to a project is responsible for the quality of the results generated from his or her task. (See Chapter 2 for more detailed information on role-related responsibilities.)

The data manager is responsible for:

- Developing protocols and SOPs to ensure data quality
- Making project leaders, technicians, etc., aware of established procedures and the importance of adhering to them
- Evaluating the quality of all data and information against NPS standards before dissemination outside the Network
- Performing periodic data audits and quality control checks to monitor and improve the data quality program

Project leaders must:

- Be aware of QA/QC protocols and convey their importance to technicians and field crews
- Ensure compliance with the protocols
- Validate data after the verification process is complete
- Review all final reports and information products

Technicians must follow established procedures for data collection, data entry, and verification included in the project protocol SOPs.

6.4 Goals and Objectives

We must ensure that a project produces data of the right type, quality, and quantity to meet project objectives and user needs. Quality criteria should be set at a level proportionate to project-specific objectives, and these criteria should indicate the level of quality acceptable for resulting data products. The EPA (2003) defines data quality objectives as qualitative and quantitative statements that:

- Clarify the intended use of the data
- Define the type of data needed to support the decision
- Identify the conditions under which the data are to be collected
- Specify tolerable limits on the probability of making a decision error due to uncertainty in the data

The most effective mechanism for ensuring that a project produces data of the right type, quality, and quantity is to provide procedures and guidelines to assist the researcher in accurate data collection, entry, and validation.

Although specific QA/QC procedures will depend upon the individual vital signs being monitored and must be specified in the protocols for each vital sign, some general concepts apply to all network projects. The general QA/QC procedures presented in this plan were primarily adapted from the Draft Data Management Protocol (Tessler & Gregson 1997) and ideas contained in Michener and Brunt (2000). These general guidelines will ensure that all data collected are checked for reliability before being integrated into the monitoring program databases. Refer to individual monitoring protocol SOPs for project-specific QA/QC requirements.

6.5 Data Collection

Careful, accurate recording of field observations in the data collection phase of a project will help reduce the incidence of invalid data in the resulting data set. Unlike a typographical error that occurs when a recorded observation is incorrectly transferred from a paper field form to a digital database, an incorrect entry in the field cannot be as easily corrected. Therefore, attention to detail during data collection is crucial to overall data quality.

Before the data collection phase of a project begins, project-specific QA/QC methods should be clearly documented. All field sheets and field data recording procedures must be reviewed and approved by the data manager and documented in the protocol. The project leader, in turn, will ensure that field crews understand the procedures and closely follow them in the field. If training is needed, the data manager will work with the project leader to provide that training. Field crew members are responsible for proofing raw data forms in the field, ensuring their readability and legibility, and verifying and explaining any unusual entries. They are expected to understand the data collection forms, know how to take measurements, and follow the established procedures.

6.5.1 Methods for Reducing Collection Errors

Use a formatted, project-specific data sheet as opposed to a field notebook. When electronic data collection devices are not used, data should be recorded on paper data forms. We strongly recommend acid-free paper to prevent fading and subsequent data loss. Some circumstances may warrant the use of paper and writing implements that can withstand moisture, dust, and other extreme environmental conditions.

Standardized data sheets that identify each piece of information to be recorded and mirror the design of the computer data entry interface will help ensure that all relevant information is recorded and subsequent data entry errors are minimized. Data sheets should contain as much basic preprinted project information as possible and sufficient space for recording relevant metadata such as date, collectors, weather conditions, etc. They should clearly specify all required information, using examples where needed to ensure that the proper values are recorded. Data recorders should adhere to the following guidelines:

- All information added to the data sheet must be printed and clearly legible.
- If alterations to the information are necessary, the original information should be crossed out with a single line and the new information written next to the original entry. Information should never be erased and old information should not be overwritten.
- Upon return from the field, copies of all original data sheets should be made and checked for legibility and completeness (*i.e.*, no data cut off at the edges). The copies of the data sheets will be stored as specified in the protocol SOP, and the original data sheets will be used for data entry.

Use a handheld computer for data collection whenever possible. The use of handheld computers eliminates the need for manual data entry from field forms and associated transcription and data entry errors. Specially designed database or computer programs may be required for handheld computers, and the user interface should be customized to the project requirements. A customized data entry application has the advantage of incorporating on-the-spot QA/QC checks, so this data collection method probably provides the highest quality data.

Use automated data loggers where appropriate. Instruments with their own data acquisition systems are useful for collecting some types of data, such as water and air quality data. These devices can be

calibrated and programmed to automatically record data and store them for later download directly to a computer, thereby eliminating the possibility for manual data entry errors.

Consider calibration, maintenance, and minimum timing requirements of field equipment. Accurate field measurements are possible only if field equipment is regularly calibrated and maintained. Where appropriate, consult SOPs and reference manuals for recommended calibration and maintenance procedures. Once in the field, allow sufficient time for field equipment to adjust to its environment so it will record accurate measurements (for example, when using water quality probes and GPS units). Researchers should maintain records of equipment calibration and failures to supplement the field data.

Be organized and keep a log. Organization is the key to good data collection methods. Maintaining a log of important decisions and events will help clarify information and contribute to an accurate report.

Ensure that field crews receive proper training. Although protocols and SOPs are in place, they alone cannot guarantee that high-quality data will be collected. Prior to routine data collection for a project, conduct training sessions to ensure that field personnel have a clear understanding of data collection procedures described in the SOPs. A training program may also include a process to certify that field staff understand and can perform the specified data collection procedures. The development of a training manual may be helpful for long-term monitoring data collection efforts and those that will involve a large number of field staff. Palmer and Landis (2002) provide an outline for a training manual and suggestions for planning training sessions.

Perform quantitative assessments of data quality. Repeating measurements is the primary tool for performing quantitative assessments of data. Project leaders should periodically review the work of field crew members to ensure that their work does not drift from standards during the course of the field season. Quantitative assessments may be considered if staff and funding are available; Palmer and Landis (2002) describe several approaches that can be used.

6.6 Data Entry

Data entry is the initial set of operations in which we transfer raw data from paper field forms into a computerized form linked to database tables. Spreadsheets should not be used for data entry (data can be exported to a spreadsheet for manipulations post entry). When data are gathered or stored digitally in the field (*e.g.*, on a data logger), data entry consists of the transfer of data (downloading) to a file on an office computer where they can be further manipulated.

Transferring data from field projects into the computer appears to be a fairly straightforward task. But the value of the data depends upon their accuracy, and we must feel confident about overall data quality. Without proper preparation and some established guidelines, data quality and integrity can be questionable. Ideally, data entry occurs as soon as possible – immediately after data collection is completed or as an ongoing process during long projects – by a person who is familiar with the data. The primary goal of data entry is *to transcribe the data from paper records into the computer with 100% accuracy*. Yet, we know that a few transcription errors are unavoidable during data entry. Thus, all data should be checked and corrected during a data verification process.

The data manager, along with the project leader, should provide training in the use of the database to all data entry technicians and other users. The project leader makes certain that data entry technicians understand how to enter data and follow the protocols. Data entry technicians are responsible for becoming familiar with the field data forms and differences in handwriting. They must also become familiar with the database software, database structure, and any standard codes for data entry used by the Network. At a minimum, data entry technicians should know how to open a data entry form, create a new

record, and exit the database properly. They must learn how to commit both a 'field' entry and a 'complete record' entry and to correct mistakes made while typing.

6.6.1 Methods for Reducing Data Entry Errors

Enter or download data in a timely manner. All data should be entered or downloaded into the project database as soon as possible, preferably at least once a week. Try to avoid delaying data entry until all the project data have been collected. Downloaded data should be periodically backed up on CD or some other semi-permanent media.

Design efficient data entry forms and methods. A full-screen data entry form that mirrors the field data form can effectively reduce manual data entry errors due to the 1:1 correspondence of the attributes. A strategy to distinguish between validated data and newly entered data should be adopted. Data can be entered into an empty, fresh database table to avoid contaminating existing data and the new data appended to the master data only after formal verification, validation, and documentation. Alternatively, we can include validation attributes that indicate which data have been checked and validated by the project leader in the database. Regardless of strategy, we must clearly document the process for validation in the protocol data management SOP.

Build automated error checking features into the database. The most robust QA/QC measures for data entry should be built into the database design to perform automatic validation checks of data. Data entry forms reduce transcription errors through auto-filled fields, range limits, pick lists, and spelling checks. They provide controlled access to the database (*i.e.*, forms are set for data entry only, which prevents accidental deletion or alteration of existing data). They control the sequence of data entry (*i.e.*, certain fields require an entry before more information can be entered). They warn the operator when errors are made and provide an opportunity for correction before the data are committed to a file.

- *Auto-filled fields.* Whenever possible, the data in a field should be auto-filled by the computer. For example, if a location ID is composed of a park code, project code, and a unique number, those elements are automatically inserted into the location ID field, ensuring that the record always contains a unique identifier.
- *Range limits.* Where the appropriate values for a particular field span a finite range, the data entry program can check the entered value against the specified minimum and maximum values for that parameter. When a value is outside the accepted range, a warning message asks the user to re-enter a valid value. For some fields, values outside a specified 'normal' range may be acceptable. In this case, the warning message asks the user to verify the entry before continuing.
- *Pick lists.* The data entry application may also use pop-up pick lists for standardized text items where spelling errors can occur. For example, rather than typing in a species code or name (where a misspelling generates a new species in the database), the code or name is selected from a list of valid species codes or scientific names and automatically entered into the species field. A pick list may also be used when only certain entries are acceptable. Lists are not suitable for all written fields but should be used when appropriate.
- *Unique constraints.* Duplicate and incorrect data entry can often be caught with the application of unique constraints on data entry fields. These constraints are particularly useful when importing data from other applications.

Provide a clean, organized work environment. Desktop space near the computer should be free of clutter and distractions that could cause the technicians to lose their place. There should be enough space for two stacks of paper documents, one from which data are being entered and one from which data have been entered. A pad or notebook and some fine colored markers should also be available for making notes. (The need for a clean workspace also applies to the verification and validation phases.)

If possible, use two data entry technicians for data entry. When one technician reads the data from the field data forms and another enters them into the computer, the work is often faster and results in a lower error rate. If only one person is available, he should work at a slower pace to avoid errors.

Perform initial and interval testing of data quality. To help ensure consistent, useful data are collected for a given monitoring objective it is important to test the data collection procedures and quality control methods soon after field work begins. A mandatory trial period (from one day to two weeks) follows thorough training for personnel involved in data collection. Data from the trial are measured against quality requirements. If the data meet the overall and protocol-specific requirements, then data collection will continue. If the data quality does not meet requirements, then personnel receive additional training and the trial is repeated followed by another quality test. If the data quality does not improve to meet requirements following the second trial, other aspects of the protocol will be examined for factors contributing to the difficulty of meeting data quality requirements.

6.7 Verification and Validation Procedures

We appraise data quality by applying verification and validation procedures as part of the quality control process. *Data verification* checks that the digitized data match the source data, while *data validation* checks that the data make sense. It is essential that we validate all data as accurate and do not misrepresent the circumstances and limitations of their collection. Failure to follow SOPs for data entry, verification, and validation will render a data set suspect. Although data entry and data verification can be handled by personnel who are less familiar with the data, validation requires in-depth knowledge about the data.

Verification and validation SOPs should be written before any project data are acquired. Technicians will follow the SOPs for verification of data, make necessary edits, and document those revisions. The project leader or designee will validate the data after verification is complete. The data manager and project leader will evaluate the results of verification and validation and determine any procedural or data form revisions that may be indicated by the results. The project leader is then responsible for reviewing all data products and reports before they are released outside the Network.

6.7.1 Methods for Data Verification

Data verification immediately follows data entry and involves checking the accuracy of the computerized records against the original source, usually hard copy field records, and identifying and correcting any errors. When we have verified the computerized data as accurately reflecting the original field data, we can archive the original paper forms and manipulate and analyze most data on the computer.

Each of the following methods has a direct correlation between effectiveness and effort. The methods that eliminate the most errors can be very time consuming while the simplest and cheapest methods will not be as efficient at detecting errors.

Visual review at data entry. The data entry technician verifies each record after input by comparing the values recorded in the database with the original values from the hard copy and immediately correcting any errors. This method is the least complicated since it requires no additional personnel or software. However, since its reliability depends entirely upon the person entering the data, it is probably the least reliable data verification method.

Visual review after data entry. All records are printed upon the completion of data entry. The values on the printout are compared with the original values from the hard copy. Errors are marked and corrected in

a timely manner. This review should be performed by someone other than the person who originally entered the data. Alternatively, two technicians can perform this review. One technician (not the one who entered the data) reads the original data sheets (the reader), and the second verifies that the same values are on the printout (the checker).

Duplicate data entry. The data entry technician completes all data entry, as normal. Random records are selected (every **n**th record) and entered into an empty replica of the permanent database by someone other than the person who entered the permanent data. We use a query to automatically compare the duplicate records from the two data sets and report any mismatches of data. Then we manually review any disparities and correct if necessary. This method adds the overhead of retyping the selected records, as well as the creation of a comparison query, but it becomes increasingly successful as the value of **n** decreases. Professional data entry services frequently use this method.

Simple summary statistics. Summary statistics can help us catch a duplicate or omitted entry. For example, we can view the number of known constant elements, such as the number of sampling sites, plots per site, or dates per sample. We can pose the same question in different ways; differences in the answer provide clues to errors. The more checks we devise to test the completeness of the data, the greater our confidence that we have completely verified the data.

To minimize transcription errors, our policy is to verify 100% of records to their original source by permanent staff. In addition, 10% of records are reviewed a second time by the project leader, and we report the results of that comparison with the data. If the project leader finds errors in this review, then we verify the entire data set again.

6.7.2 Methods for Data Validation

Although we may have correctly transcribed the data from original field notes or forms, they may still be inaccurate or illogical. For example, an entry of stream pH of 25.0 or a temperature of 95°C in data files is undoubtedly incorrect, even if it was correctly transcribed from the field form. Validation can accompany data verification *only* if the operator has comprehensive knowledge about the data. More often, validation is a separate operation carried out *after* verification by the project leader or resource specialist who can identify generic and specific errors in particular data types. Corrections or deletions of logical or range errors in a data set require notations on the original paper field records about how and why the data were changed. Modifications of the field data should be clear and concise while preserving the original data entries or notes (*i.e.*, no erasing!). Validation should also include a completeness check of the data set since field sheets or other sources of data could easily be overlooked.

General step-by-step instructions are not possible for data validation because each data set has unique measurement ranges, sampling precision, and accuracy. Nevertheless, validation is a critically important step in the certification of the data. Invalid data commonly consist of slightly misspelled species names or site codes, the wrong date, or out-of-range errors in parameters with well defined limits (*e.g.*, elevation). But more interesting and often puzzling errors are detected as unreasonable metrics (*e.g.*, stream temperature of 70°C) or impossible associations (*e.g.*, a tree 2 feet in diameter but only 3 feet high). We call these types of erroneous data *logic errors* because using them produces illogical (and incorrect) results. The discovery of logic errors has direct, positive consequences for data quality and provides important feedback to the methods and data forms used in the field. Histograms, line plots, and basic statistics can reveal possible logic and range errors.

The following general methods can be used to validate data. Specific procedures for data validation depend upon the vital sign being monitored and will be included in the monitoring protocol SOPs.

Data entry application programming. Certain components of data validation are built into data entry forms. The simplest validation during data entry is range checking, such as ensuring that a user attempting to enter a pH of 20.0 gets a warning and the opportunity to enter a correct value between 1.0 and 14.0 (or better yet, within a narrow range appropriate to the study area). Not all fields, however, have appropriate ranges that are known in advance, so knowledge of what are reasonable data and a separate, interactive validation stage are important.

Edwards (2000) suggests the use of 'illegal data' filters, which check a specified list of variable value constraints on the master data set (or on an update to be added to the master) and create an output data set. This output data set includes an entry for each violation, along with identifying information and an explanation of the violation. He illustrates the structure of such a program, written in the SAS® programming language.

A caveat should be interjected regarding the operative word 'illegal.' Even though a value above or below a given threshold has never before been observed and the possibility that it could occur seems impossible, such an observation is not always an illegal data point. Edwards (2000) points out that one of the most famous data QA/QC blunders to date occurred when NASA's computer programs deleted satellite observations of ozone concentrations that were below a specified level, seriously delaying the discovery of the ozone hole over the South Pole.

Outlier Detection. According to Edwards (2000), "the term outlier is not (and should not be) formally defined. An outlier is simply an unusually extreme value for a variable, given the statistical model in use." Any data set will undoubtedly contain some extreme values, so the meaning of 'unusually extreme' is subjective. The challenge in detecting outliers is in deciding how unusual a value must be before it can (with confidence) be considered 'unusually' extreme.

Data quality assurance procedures should not try to **eliminate** outliers. Extreme values naturally occur in many ecological phenomena; eliminating these values simply because they are extreme is equivalent to pretending the phenomenon is 'well-behaved' when it is not. Eliminating data contamination is a better way to explain this quality assurance goal. If contamination is not detected during data collection, it is usually only detected later if an outlying data value results. When we detect an outlier, we should try to determine if some contamination is responsible.

We can use database, graphic, and statistical tools for ad-hoc queries and displays of the data to detect outliers. Some of these outlying values may appear unusual but prove to be quite valid after confirmation. Noting correct but unusual values in documentation of the data set saves other users from checking the same unusual values.

Other exploratory data analyses. Palmer and Landis (2002) suggest that, in some cases, calculations for assessments of precision, bias, representativeness, completeness, and comparability may be applicable. For certain types of measurements, evaluation of a detection limit may also be warranted (the authors provide examples of procedures that may be applicable). Normal probability plots, Grubb's test, and simple and multiple linear regression techniques may also be used (Edwards, 2000; the author provides SAS and Splus code for constructing normal probability plots and examples of output showing normal and non-normal distributions).

6.8 Version Control

The Network manages files from a multitude of sources, comprised of many formats with many iterations of a particular product. Some files are complete, others are works-in-progress, and the status of some

may not be known. Determining the status of a single file can be difficult, but determining the current file within a series of similarly named files can be almost impossible.

Version control is the process of documenting the temporal integrity of files as they are being changed or updated. Change includes any alteration in the structure or content of the files, and such changes should not be made without the ability to undo mistakes caused by incorrect manipulation of the data. Whenever we complete a set of data changes, we should save the file with a unique name, a simple act that should become routine for all data handlers.

Several options for version control may be used:

- *Dates.* Using a date provides logical version control. The date is usually formatted as YYYYMMDD or YYMMDD, where DD is optional (depending on the frequency of changes).
- *Sequential numbers.* We can designate versioning of archived data sets by adding a number to the file name; *e.g.*, 001 or V1.0 for the first version; each additional version receives a sequentially higher number. We should also document the date that a file becomes a new version, perhaps through assignment of database folder (or directory) names.
- *Version control software.* We can eliminate the work of differentiating multiple versions of documents by using version control software to append modifying characters to the file name. Such software applications track changes made to a document, add comments related to the different document iterations, and retrieve the document at any recorded stage of development. These applications are available in either desktop or online formats.
- *Database software*. Fields to record the file's revision history may be incorporated into databases. Built-in backup routines that allow for automatic file renaming and archiving may also be considered.

Prior to any major changes to a file, we should store a copy of the file with the appropriate version number to allow the tracking of changes over time. With proper controls and communication, versioning ensures that only the most current version is used in any analysis.

Currently, the Network uses dates for version control. The most recent or final version of a file is stored without a date indicator; previous versions have a date appended to the file name. Other methods, including version control software, will be evaluated for adoption.

6.9 Data Quality Review and Communication

The National Park Service requires QA/QC review prior to communicating/disseminating data and information, and only data and information that adhere to NPS quality standards can be released. Data are distributed to the public through the network webpage, national websites such as the Biodiversity Data Store and the NR-GIS Data Store, and public access databases such as NPSpecies and NatureBib. Information distributed through any of these mechanisms must undergo internal QA/QC procedures, and the standards used in producing the information and that substantiate its quality must be formally documented. Information disseminated to the public must be approved by the appropriate reviewing officials and programs. Mechanisms for receiving and addressing comments/complaints pertaining to the quality of data must also be in place.

6.9.1 Data Quality Review Methods

The SODN I&M Program will establish guidelines to ensure compliance with DO #11B. This guidance will document both internal and external review procedures for data and information disseminated outside the Network, as well as a process for processing complaints about data quality.

Edwards (2000) suggests the initiation of quality circles, regular meetings of project leaders, the data manager, and data management personnel for discussing data quality problems and issues. These meetings promote teamwork attitudes while focusing brainpower on data quality issues. Participants become more aware of quality issues and learn to anticipate problems. Moreover, all participants develop a greater appreciation of the importance of their role in data quality for the entire monitoring effort.

6.9.2 Value of Feedback from QA/QC Procedures

Quality assurance procedures may require revision if random checks reveal an unacceptable level of data quality. However, quality checks should not be performed with the sole objective of eliminating errors, as the results may also prove useful in improving the overall process. For example, if the month and day are repeatedly reversed in a date field, the data entry technicians may require retraining about the month/day entry order. If retraining is unsuccessful in reducing the error's occurrence, the digital data entry form may need to be revised so that month and day are entered separately, field length limits are enforced, or a pick list is created. In this manner, the validation process will serve as a means of improving quality as well as controlling the lack of quality.

Following validation, re-evaluation of the field data forms as the source of logic errors is recommended, and we can modify the forms to avoid common mistakes if necessary. Often minor changes, small annotations, or the addition of check boxes to a form can remove ambiguity about what should be entered. Perhaps surprisingly, when we find the same type of validation error occurring repeatedly in different data sets, the field form – not the field crew – is usually at fault. Repeated errors found during validation can also mean that protocols or field training are at fault, so these should be reviewed and corrected as needed.

6.9.3 Monitoring Conformance to Standards and Protocols

Data managers can use periodic data audits and quality control inspections to maintain and improve their data quality procedures. They must verify that staff is operating in conformance with the data quality procedures specified in this plan and the protocol-specific data management SOPs. They should track and facilitate the correction of any deficiencies. These quality checks promote a cyclic process of continuous feedback and improvement of both the data and the quality planning process.

Periodic checks by the data manager to see if network staff are adhering to the data quality procedures established in the Data Management Plan and protocol SOPs may include verification of the following:

- Data collection and reporting requirements are being met
- Data collection and reporting procedures are being followed
- Verification and validation procedures are being followed
- Data file structures and maintenance are clear, accurate, and according to plan
- Revision control of program documents and field sheets are adequate
- Calibration and maintenance procedures are being followed
- Seasonal and temporary staff are trained in data management practices
- Metadata collection and creation for the program are proceeding in a timely manner
- Data are being archived and cataloged appropriately for long-term storage

The results of quality assessments should be documented and reported to the project leader and the network coordinator. The project leader and coordinator are responsible for ensuring that non-conformities in data management practices are corrected.

6.9.4 Communicating Data Quality

The Network will use data documentation and metadata to notify end users, project leaders, and network management of data quality (see Chapter 7). A descriptive document for each data set/database will provide a 'quality report' (*i.e.*, information on the specific QA/QC procedures applied and the results of the review). Descriptive documents and/or formal FGDC-compliant metadata will document quality for spatial and non-spatial data files posted on the Internet.

Credits

This chapter was developed from material in the cited references and suggestions/comments from Teresa Leibfreid (Cumberland Piedmont Network), Bill Moore (Mammoth Cave National Park), Velma Potash (Cape Cod National Seashore), Geoffrey Sanders (National Capital Region Network), and Brian Witcher (San Francisco Bay Area Network).

Chapter 7. Data Documentation

Thorough, complete, and accurate documentation is critical during every stage of processing in the life cycle of a data set. At times, data sets appear to take on lives of their own. Some seem to have the ability to reproduce and evolve on multiple hard drives, servers, and other storage media. Others are masters at remaining hidden in outdated digital formats or in forgotten file drawers. In addition, once data are discovered, a potential user is often left with little or no information regarding the quality, completeness, or manipulations performed on a particular 'copy' of a data set. Such ambiguity results in lost productivity as the user must invest time in tracking down information or, worst case scenario, renders the data set useless because answers to these and other critical questions cannot be found. Therefore, data documentation must include an upfront investment in planning and organization.

This chapter will focus on data documentation as a critical step toward ensuring that all data sets retain their integrity and utility well into the future. Complete, thorough, and accurate documentation should be of the highest priority for long-term studies, and since long-term data sets are continually changing, this documentation must remain up-to-date. Data documentation involves the development of metadata, which at the most basic level can be defined as 'data about data,' or more specifically as information about the content, context, structure, quality, and other characteristics of a data set. Metadata provide the information necessary to relate the raw data to the underlying theoretical or conceptual model(s) for appropriate use and interpretation (Michener 2000). Additionally, standardized metadata provide a means to catalog data sets within Intranet and Internet systems, thus making these data sets available to a broad range of potential users.

Past efforts to standardize metadata content and format have focused primarily on geospatial data sets. Therefore, the term 'metadata' is often associated with documentation compliant with formal Federal Geographic Data Committee (FGDC) standards. However, in this plan, the term 'metadata' encompasses all forms of data documentation, including those for spatial and non-spatial tabular data, as well as project-level documentation.

7.1 Data Set Documentation

While the importance for metadata is universally accepted within the data management community, the approaches for collection and levels of detail are varied (sometimes referred to as the '101 ways'). However, we should consider the following when developing data documentation strategies.

- Executive Order 12906, signed by President William Jefferson Clinton in 1994, mandates each federal agency to "...document all new geospatial data it collects or produces, either directly or indirectly..." using the Federal Geographic Data Committee (FGDC) Content Standard for Digital Geospatial Metadata (CSDGM). In addition, EO 12906 directs agencies to plan for legacy data documentation and provide metadata and data to the public. (See the CSDGM website at http://www.fgdc.gov/metadata/contstan.html for more information.)
- The FGDC Biological Data Profile (http://www.fgdc.gov/standards/status/sub5_2.html) contains all the elements of the CSDGM and includes additional elements for describing biological data sets with a spatial component. Metadata created in compliance with the Biological Data Profile can be added to the National Biological Information Infrastructure (NBII) Clearinghouse (http://www.nbii.gov/datainfo/metadata/). Although not a requirement, we strongly recommend completion of the Biological Data Profile for appropriate data sets.
- All GIS data layers must be documented with applicable FGDC and NPS metadata standards. The NPS GIS Committee requires all GIS data layers be described with FGDC standards and the NPS Metadata Profile (http://nrdata.nps.gov/profiles/NPS_Profile.xml).

- While there are numerous tools available for developing metadata, the NPS Integrated Metadata System Plan (http://science.nature.nps.gov/im/datamgmt/metaplan.htm) is limited to three recommended desktop applications: Dataset Catalog, ArcCatalog, and Spatial Metadata Management System (SMMS).

7.1.1 NPS Integrated Metadata System Plan and Tools

As noted above, the NPS Integrated Metadata System Plan is limited to three recommended desktop applications for collecting metadata. These include Dataset Catalog (developed by the I&M Program) and two commercial off-the-shelf metadata tools, ArcCatalog and SMMS. A brief description of each of these tools, including its potential utility in metadata creation follows. A fourth tool, the Metadata Parser (mp), and its utility in metadata creation is also briefly discussed.

Dataset Catalog

The Dataset Catalog (http://science.nature.nps.gov/im/apps/datacat/index.htm) is a tool for cataloging abbreviated metadata on geospatial and biological data sets pertaining to park(s) and/or a network. It provides parks and/or networks a tool to inventory, organize, and maintain information about data set holdings locally. While Dataset Catalog is not intended to be an exhaustive metadata listing, it does assist parks and networks in beginning to meet the mandates of EO 12906. The current version of Dataset Catalog (version 3) can export records as FGDC text files and in Extensible Markup Language (XML), either of which can then be imported into other metadata tools. The I&M Program recommends that all relevant data sets at I&M parks and networks be cataloged in at least simple Dataset Catalog format, and the Network will follow this recommendation.

Spatial Metadata Management System

The Spatial Metadata Management System (SMMS; http://imgs.intergraph.com/smms/) is a tool used to create, edit, view, and publish metadata that is compliant with FGDC requirements. SMMS uses an MS Access database structure combined with an advanced FGDC-compliant metadata editor. The software allows selection of views depending on whether the user wants the full standard, biological, or the minimal compliant view of Sections 1 and 7. Online Help describes the purpose, usage, or mandatory status of metadata elements. The context-sensitive help file provides the FGDC definition for each field on the screen. In addition to Help files, there are sample metadata records for most sections that provide 'real world' examples. The NPS Integrated Metadata System Plan recommends SMMS for FGDC Biological Profile and other geospatial metadata creation.

ArcCatalog

ArcCatalog (http://www.esri.com/software/arcgis/arcinfo/index.html) is a management tool for GIS files contained within the ArcGIS Desktop suite of applications. With ArcCatalog, users can browse, manage, create, and organize tabular and GIS data. In addition, ArcCatalog comes with support for several popular metadata standards that allow one to create, edit, and view information about the data. There are editors to enter metadata, a storage schema, and property sheets to view the data. Users can view GIS data holdings, preview geographic information, view and edit metadata, work with tables, and define the schema structure for GIS data layers. Metadata within ArcCatalog are stored exclusively as Extensible Markup Language (XML) files. The NPS Integrated Metadata System Plan recommends ArcCatalog for gathering GIS-integrated geospatial metadata. An optional but highly recommended extension for ArcCatalog is the NPS Metadata Tools & Editor, which incorporates the functions previously available in the NPS Metadata ArcCatalog Extension developed by NPS Midwest Region GIS Technical Support Center. The NPS Metadata Tools & Editor is also available as a standalone desktop interface. The Tools & Editor is intended to be the primary editor for metadata that will be uploaded to the NR-GIS Data Store, and the Network is currently using ArcCatalog and this tool to create such metadata.

Metadata Parser

The Metadata Parser (mp; http://geology.usgs.gov/tools/metadata/tools/doc/mp.html) program is used to validate metadata records by checking the syntax against the CSDGM. The program generates a textual report indicating errors in the metadata, primarily in the structure but also in the values of some of the scalar elements where values are restricted by the standard. It also generates FGDC-compliant output files for posting to spatial data clearinghouses. The Metadata Parser is now bundled with the NPS Metadata Tools & Editor (Parse with MP tool), and the Network uses mp to check all metadata files destined for the NR-GIS Data Store.

The NPS approach to unify and streamline metadata development (Figure 7.1) utilizes existing desktop metadata creation applications, as well as an online integrated metadata database (the NR-GIS Metadata Database) and a web based data server (NR-GIS Data Server). The NR-GIS Metadata and Data Store (http://science.nature.nps.gov/nrdata/) will comprise a web-based system to integrate data dissemination and metadata maintenance. Records may be updated in the NR-GIS Metadata Database (Dataset Catalog only) or in the source desktop application (*i.e.*, ArcCatalog, Dataset Catalog, SMMS). Non-sensitive NR-GIS Metadata records are automatically posted to NPS Focus (http://focus.nps.gov/).

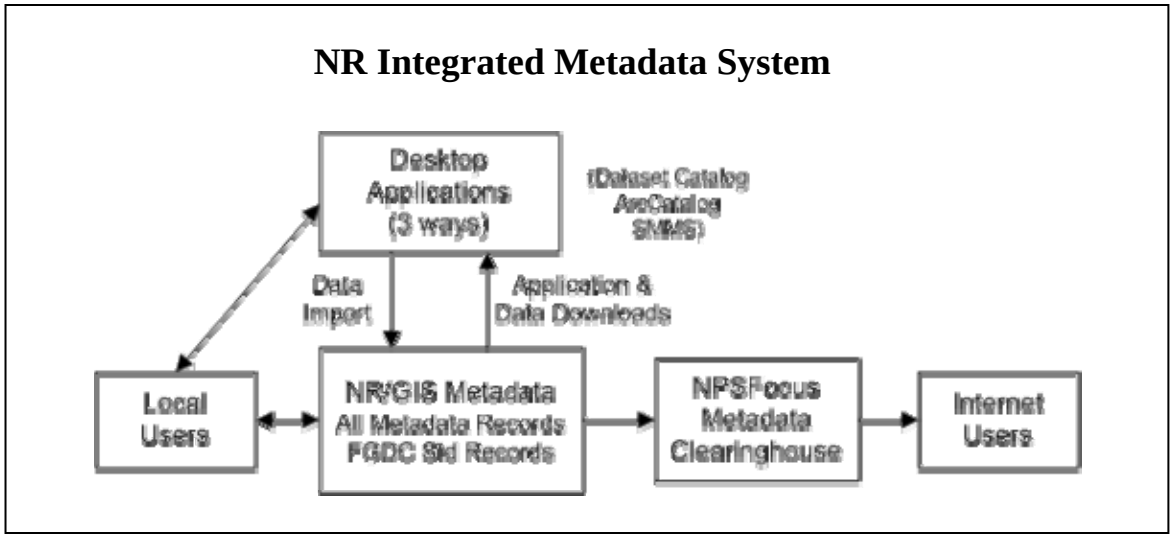

Figure 7.1. Natural Resources Integrated Metadata System (http://science.nature.nps.gov/im/datamgmt/metaplan.htm).

7.1.2 Metadata Process/Work Flow

The general process for creating formal metadata records is described below.

Identify Relevant Data Sets and Compile Pertinent Metadata. Data utilized by the Network can be grouped, at least initially, into three broad categories based on origin. These categories include programmatic data sets, other NPS data sets (often legacy data), and external data sets. (See Chapter 5 for more information on these categories of data.)

- *Programmatic Data* – metadata development will begin up front for new projects by providing the project leader with a copy of the Network's data documentation standards. In most instances, this will include completion of a basic metadata survey for inclusion in the data manager's project file, as well as submission of supporting documentation (proposal, SOPs, etc.). A processing and

revision log will be maintained with the data set to document all original processing steps and subsequent revisions or updates to the data. Updates and revisions to the metadata for long-term data sets will be conducted in tandem with data submissions.

- *Other NPS Data* – in many cases, legacy data are initially identified through data mining efforts. Unfortunately, many legacy data sets will be missing pertinent information, and the originator may no longer be available for consultation. Thus, an 'adequate' level of documentation may not be possible. Nevertheless, the data and all related supporting documentation should be assembled and reviewed. (See Chapter 5 for further information on legacy data acquisition and processing.) Many legacy tabular data sets will require conversion to a standard database format for integration with other data for future analyses. Data entry, verification, and validation procedures will follow those contained within this plan (Chapter 6). A processing and revision log will be maintained with the data set to document all original processing steps and subsequent revisions or updates to the data.
- *External Data* – networks are not the only entities gathering natural resource data relevant to park management, and we will make every effort to capture and assimilate all relevant data. Outside entities will be contacted and requests made for available metadata, and/or a metadata interview will be conducted. As with legacy data, tabular data files may require conversion to a standard database format for analysis. Data entry, verification, and validation procedures will follow those contained in Chapter 6 of this plan and specific monitoring protocols. A processing and revision log will be maintained with the data set to document all original processing steps and subsequent revisions or updates to the data.

Create Dataset Catalog Record. Because metadata documentation can be accomplished in a variety of formats and levels of detail, it is sometimes viewed as an overwhelming task. As a starting point, the Network will develop a simple Dataset Catalog record for relevant spatial and non-spatial data. This approach provides brief metadata for all network data holdings in a searchable, centralized location. In addition, managers can identify and prioritize data sets for which formal metadata will be developed and identify the status of metadata documentation for a particular data set (*e.g.*, planned, in work, complete). These records can be imported into the online NR-GIS Metadata Database or continue through additional processing steps based on data type, source, and importance.

Prioritization of legacy data sets for further documentation will be based upon current or anticipated future use. In other words data sets with a high likelihood of repeated use in analyses or those with a high probability for data sharing will be addressed first. All GIS layers will be documented with applicable FGDC and NPS metadata standards. In addition, we strongly recommend that all data sets/databases be described in minimal to complete NBII Biological Profile FGDC metadata format, as appropriate. At a minimum, the data set will be entered into Dataset Catalog.

Create FGDC-Compliant Metadata Record. Where appropriate, use ArcCatalog and the NR-GIS Metadata Tools & Editor to complete FGDC-compliant metadata. Parse the metadata record with the Metadata Parser.

Make Information Available. At a minimum, metadata and associated non-sensitive data will be submitted to the NR-GIS Metadata and Data Store. Additionally, information on data holdings should be conveyed in a meaningful manner to park resource managers, researchers, and others with a potential interest/stake in park management and/or research endeavors. Similar to metadata creation the mechanisms and formats for accomplishing this are varied. In addition to FGDC text and XML files, Dataset Catalog can output a list of all records, single record reports, and/or a data dictionary report. Dependent upon the target audience, these standardized outputs can be useful in conveying information on Program data holdings and summaries of database structures. In addition, customized queries and

reports can also be generated. Other standardized outputs include ArcCatalog stylesheets. The NPS Metadata Tools & Editor contains custom stylesheets, which can be invoked from the metadata toolbar. These can be utilized to depict pertinent details in a more coherent format than standard metadata outputs.

7.1.3 Additional Documentation for Tabular Data Sets

Proper use and interpretation of the raw data contained in the rows and columns (observations and fields) of a tabular data file depend upon the context supplied by the accompanying documentation. Databases delivered to the Network for all funded projects will be accompanied by a descriptive document that includes the following information about the project and the data:

- Contents of the CD or zip file containing the data set
- Description of the project
- Location of the project study plan and work plan
- Project leader's name and contact information
- Principal investigator's name and contact information
- Data set contact's name and contact information
- Description of the database model (entity-relationship diagram and data dictionary)
- Sensitive data issues, if appropriate
- Description of data verification/validation methods and results (data quality report)
- Certification of the data set
- Additional comments/documentation references, where appropriate

Documentation for data sets created by the Network will include:

- Description of the database model
- Entity-relationship diagram
- Data dictionary (use the Harvester in Dataset Catalog)
- Data quality report
- Sensitive data report
- Certification of the data set

7.2 Project Documentation

The metadata for each project will include a digital project index document that links to various project-related products, such as data sets, reports, maps, etc.

The project tracking application will contain a complete list of expected products. Once they have been delivered, this function will serve as a reference for available products.

Credits

The majority of this chapter was adapted from material developed by Teresa Leibfreid (Cumberland Piedmont Network) and Bill Moore (Mammoth Cave National Park).

Chapter 8. Data Management Support for Analysis and Reporting

The success of the SODN I&M Program depends upon providing the information park managers need to make science-based decisions on the management of natural resources in their park, as well as disseminating this information to a wider audience consisting of other NPS staff, other agency personnel, external scientists, and the general public. Data analyses are the means by which we transform data into this essential information. In the case of inventories, short-term, or special projects, data may be analyzed only once, at the project's completion. For long-term monitoring studies, data should be summarized at least annually and fully analyzed at three-to-five-year intervals (or as specified in the monitoring protocols) in order to detect trends in resource conditions. The information derived from data analyses will be conveyed through a variety of written reports and presentations. Project leaders are ultimately responsible for analyzing data and reporting the results, but this chapter discusses how data management support can facilitate those activities through automated data summaries and reports.

8.1 Timeline for Analysis and Reporting

Each project will have a schedule for data analysis and reporting requirements specified in the monitoring protocol, study plan, cooperative agreement, or contract. However, in general, the Network will complete data analysis and reporting within one year of seasonal data collection or the end of the project. Each project leader will maintain his or her working data set throughout the duration of the project. Once a year for long-term projects or within a year of completion of a short-term project, he or she is responsible for reviewing and certifying the data set, producing appropriate data summaries/analyses, writing a report, and submitting the data set to the data manager, who will place it in the appropriate repository and/or integrate it into the master database, where it will be available for other syntheses or analyses.

Specific timelines for data processing will be established for each monitoring vital sign. Chapter 7 of the Sonoran Desert Network Vital Signs Monitoring Plan describes the reports that the Network will produce.

8.2 Coordination with Project Leaders

Close coordination between the project leader and the data manager is important when defining the process of transforming project data from raw data into meaningful information. Based on project objectives, protocols, and data management and analysis SOPs, the project leader and data manager will work together to identify opportunities and methods to streamline data extraction and exports from databases and to automate summaries, analyses, and reports.

8.2.1 Data Extraction

Extracting the data needed for a summary or analysis requires an understanding of the database design, including the relationships (visualized in an ERD, or Entity-Relationship Diagram), table structures, and fields within each table. In many cases, the pieces of information necessary for a summary or analysis will come from more than one table. After identifying the required fields, MS Access queries may be written to extract data from multiple tables. The project leader and data manager will work together to create and test these queries and evaluate the resulting products.

8.2.2 Data Exports

Basic statistics such as means, standard deviations, and other descriptive statistics may be calculated within an MS Access database. However, in most cases, data from MS Access will be exported in other formats for further analysis in more robust statistical programs. The preferred alternative is a file in .dbf

format exported directly from the Access database. In cases where this format is not appropriate, ASCII text files have the advantage of being almost universally readable by third party applications, although ASCII text requires extra steps to transfer data between applications. Data may be exported to MS Excel spreadsheets for calculating very basic descriptive statistics, but Excel should not be used for most statistical analyses. Third-party applications, such as SAS and Primer, will be used for procedures such as frequency distribution plots; tests for normality; single, multivariate, and repeated measures analysis of variance; and regression analyses.

The project leader and data manager will review the objectives of the analysis prior to exporting data. The order of variables in the resulting 'flat file' is most easily controlled through query design in the relational database, so up-front planning is recommended. Changing data types should not be required if the database was properly designed at the outset of a project. If a change is required, care must be taken to minimize the risk of data loss.

8.2.3 Automated Data Summaries and Reports

Where possible, the Network will use built-in functions or custom programming in project databases to facilitate the production of data summaries and reports. The project leader will determine the necessary outputs based on project objectives, protocols, and reporting requirements. He or she will discuss these specifications with the data manager, who will develop the queries, macros, and any other programming to produce the desired data summaries, which can then be incorporated into automated reports.

8.3 Annual Analyses and Reports

Yearly project reports are required for all long-term projects – monitoring vital signs as well as other multi-year studies. Annual reports should convey the past year's network monitoring activities to park resource managers as well as network staff and other scientists. Relevant information may include numbers of samples for each park and relevant attributes, data management activities, any changes made to the protocols, and the status of resources. Annual reports will be written as part of the SODN Technical Report series and follow the standards for those reports. A summary of each annual report that highlights key points will also be produced in a 'brochure' format for distribution to a wider audience, including park superintendents and the general public.

Annual 'state of the parks' reports describing the current trends and conditions of park resources will also be prepared for use at the national level to prepare summaries for Congress, NPS leadership, park superintendents, and the general public.

Summaries of work completed on national databases will be included in the Annual Administrative Report and Work Plan (AARWP). These summaries will provide information such as number of records added for NPSpecies, NatureBib, and the NR-GIS Metadata Database and number of products added to the NR-GIS Data Server and the Biodiversity Data Store.

8.4 Long-Term Analyses and Reports

Comprehensive reports incorporating detailed data analyses, syntheses, and descriptions of trends in resource conditions for each vital sign will be produced every 3-5 years or according to the individual monitoring protocol requirements. These reports will also be produced as part of the SODN Technical Report series according to the standards for those reports. Park-specific reports will describe and interpret trends in the monitored resource and the relationships among resources and recommend alternatives for resources in need of management action. Network-level reports will describe the role of environmental differences in resource trends among parks and interpret these trends in Network and regional contexts.

Similar to annual reports, summaries of these reports highlighting key findings and recommendations will be produced as 'brochures' targeted to park superintendents, interpretation staff, and the general public.

8.5 Special Analyses and Reports

Biological Inventories

Biological inventories of vertebrates and vascular plants in Network parks began in 2000 and were completed in 2005. Basic summaries were produced and annual reports on the findings have been completed for the 2000, 2001, and 2002 field seasons. A final inventory report for each park will also be completed.

Protocol Development and Pilot Projects

As procedures are tested and adopted, progress reports describing the background and the methods of protocol development and enhancements may be required on an annual basis. After a protocol has been adopted for use by the Network, a final report that documents the record of decision for protocol design and the results of pilot studies is required. Thereafter, reports will be produced on an as-needed basis (*e.g.*, upon modification of a protocol).

Credits

Portions of this chapter were adapted from material developed by Gareth Rowell and Mike Williams (Heartland Network).

Chapter 9. Data Dissemination

One of the most important goals of the Inventory and Monitoring Program is to integrate natural resource inventory and monitoring information into National Park Service planning, management, and decision making. To accomplish this goal, relevant natural resource data must be cataloged, archived, and made available to park managers. This chapter describes the steps the SODN I&M Program will take to disseminate data and information, keeping in mind NPS policies on data ownership and the protection of 'sensitive' natural resource information.

9.1 Data Ownership

The National Park Service defines conditions for the ownership and sharing of collections, data, and results based on research funded by the United States government. All cooperative and interagency agreements, as well as contracts, should include clear provisions for data ownership and sharing as defined by the National Park Service:

- All data and materials collected or generated using National Park Service personnel and funds become the property of the National Park Service.
- Any important findings from research and educational activities should be promptly submitted for publication. Authorship must accurately reflect the contributions of those involved.
- Investigators must share collections, data, results, and supporting materials with other researchers whenever possible. In exceptional cases, where collections or data are sensitive or fragile, access may be limited.

9.1.1 Office of Management and Budget Policy on Data Ownership

The Office of Management and Budget (OMB) ensures that grants and cooperative agreements are properly managed. Federal funding must be disbursed in accordance with applicable laws and regulations. OMB circulars establish some degree of standardization government-wide to achieve consistency and uniformity in the development and administration of grants and cooperative agreements. Specifically, OMB Circular A-110 establishes property standards within cooperative agreements with higher institutions and non-profit organizations. Section 36 of Circular A-110 regarding 'Intangible Property' describes the following administrative requirements pertinent to data and ownership:

> *(a) The recipient* (academic institution or non-profit organization receiving federal monies for natural resource inventory and/or monitoring) *may copyright any work that is subject to copyright and was developed, or for which ownership was purchased, under an award. The Federal awarding agency(ies*) (in this case the National Park Service) *reserve a royalty-free, nonexclusive and irrevocable right to reproduce, publish, or otherwise use the work for Federal purposes, and to authorize others to do so.*
>
> Section 36 also states:
>
> *(c) The Federal Government has the right to:*
>
> > *(1) obtain, reproduce, publish or otherwise use the data first produced under an award*
> >
> > *(2) authorize others to receive, reproduce, publish, or otherwise use such data for Federal purposes*

(d) (1) In addition, in response to a Freedom of Information Act (FOIA) request for research data relating to published research findings produced under an award that were used by the Federal Government in developing an agency action that has the force and effect of law, the Federal awarding agency shall request, and the recipient shall provide, within a reasonable time, the research data so that they can be made available to the public through the procedures established under the FOIA (5 U.S.C. 552(a)(4)(A)).

(2) The following definitions apply for purposes of paragraph (d) of this section:

(i) Research data is defined as the recorded factual material commonly accepted in the scientific community as necessary to validate research findings, but not any of the following: preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues. This "recorded" material excludes physical objects (e.g., laboratory samples).…

(ii) Published is defined as either when:

(A) Research findings are published in a peer-reviewed scientific or technical journal; or

(B) A Federal agency publicly and officially cites the research findings in support of an agency action that has the force and effect of law.

(iii) Used by the Federal Government in developing an agency action that has the force and effect of law is defined as when an agency publicly and officially cites the research findings in support of an agency action that has the force and effect of law.

9.1.2 Establishing Data Ownership Guidelines

The Sonoran Desert Network has established guidelines for the ownership of data and other research information. To ensure that proper ownership, format, and development of network products is maintained, all cooperative or interagency work must be conducted as part of a signed collaborative agreement. Every cooperative or interagency agreement or contract involving the Network must include OMB Circular A-110 cited under the *Reports and Deliverables* section of all agreements and contracts. The following shows appropriate language to use when citing Circular A-110:

> "As the performing organization of this agreement, *institution or organization name* shall follow the procedures and policies set forth in OMB Circular A-110."

Every cooperative or interagency agreement or contract must include a list of products or deliverables clearly defined within each agreement or contract. These include, but are not limited to, field notebooks, photographs (hardcopy and digital), specimens, raw data, and reports. Details on formatting and media types that will be required for final submission must be included.

The following statement must also be included in the *Reports and Deliverables* section of all Network agreements and contracts:

> "All reports and deliverables must follow the most recent version of the Sonoran Desert Network Product Specifications."

Researchers should also provide a schedule of deliverables that includes sufficient time for NPS review of draft deliverables before scheduled final submissions.

9.2 Data Distribution

One of the most important goals of the I&M Program is to *integrate natural resource inventory and monitoring information into National Park Service planning, management, and decision making.*

To accomplish this goal, procedures must be developed to ensure that relevant natural resource data collected by NPS staff, cooperators, researchers, and the public are entered, quality-checked, analyzed, documented, cataloged, archived, and made available for management decision-making, research, and education. Providing well-documented data in a timely manner to park managers is especially important to the success of the Program. The SODN I&M Program will make certain that:

- Data are easily discoverable and obtainable.
- Data that have not yet been subjected to full quality control will not be released by the Network, unless necessary in response to a FOIA request OR unless accompanied by a 'quality disclaimer' that also prohibits further release by the recipient.
- Distributed data are accompanied by complete metadata that clearly establish the data as products of the NPS I&M Program.
- Sensitive data are identified and protected from unauthorized access and inappropriate use.
- A complete record of data distribution/dissemination is maintained.

To accomplish this, the Network will use a number of distribution methods that will make information collected and developed as part of the Program widely available to park staff and the public.

9.2.1 Data Distribution Mechanisms

The Network's main mechanism for distribution of its inventory and monitoring data will be the Internet. Use of the Internet for dissemination will allow the data and information to reach a broad community of users. As part of the NPS I&M Program, web-based applications and repositories have been developed to store a variety of park natural resource information. The Network will use the following applications and repositories to distribute data developed by the program:

- *NatureBib* – master web-based database housing natural resource bibliographic data for I&M Program parks.
- *NPSpecies* – master web-based database to store, manage, and disseminate scientific information on the biodiversity of all organisms in all National Park units.
- *Biodiversity Data Store* – digital archive of documents, GIS data sets, and non-GIS data sets that document the presence/absence, distribution, and/or abundance of any taxa in National Park Service units.
- *NR-GIS Metadata and Data Store* – online repository for metadata and associated data products.
- *Sonoran Desert Network website* – detailed information about the Network and its I&M Program. Metadata on all inventory and monitoring products developed as part of the Network's I&M plan will be posted to this site. Data and products will either be available through the site, or users will be directed to where the data are stored.

The table (9.1) below provides a list of data types that will be uploaded to these sites.

Table 9.1. Types of data uploaded to web applications/repositories.

Web Application Name	Data Types Available at Site
NPSpecies	Data on park biodiversity (species information)
NatureBib	Park-related scientific citations
Biodiversity Data Store	Raw or manipulated data and products associated with inventory and monitoring data that have been entered into NPSpecies
NR-GIS Metadata and Data Store	Spatial and non-spatial metadata and data products
SODN Website	Reports and metadata for all inventory and monitoring data produced by the Network

Because network data will reside in the repositories listed above, these data will automatically be searchable via the integrated metadata and image management system and search gateway called NPS Focus. This system is being built with Blue Angel Enterprise software for metadata management and the LizardTech Express Server for image management. Currently ten NPS and two non-NPS databases have been integrated into the NPS Focus prototype in either full or testbed form for one-stop searching. NPS Focus has been released as an Intranet version only (http://focus.nps.gov/); a public version is projected in the near future.

Of the ten databases uploaded to NPS Focus to date, NatureBib and the NR-GIS Metadata and Data Store are most applicable to existing network bibliographic and spatial metadata. The Network will continue to upload data and information to these two sites, which will coincide with the ability to search for these records through the NPS Focus portal. As NPS Focus reaches further development, other databases and repositories utilized by the Network are expected to be searchable through this portal as well.

9.2.2 Data Classification: Protected vs. Public

All data and associated information from I&M activities must be assessed to determine their sensitivity. This includes, but is not limited to, reports, metadata, raw and manipulated spatial and non-spatial data, maps, etc. Network staff must carefully identify and manage any information that is considered sensitive. The Network must clearly identify and define those data needing access restrictions and those to make public.

The Federal Geographic Data Committee's Homeland Security Working Group (http://www.fgdc.gov/participation/working-groups-subcommittees/hswg/index_html) has published an interim version of the "Guidelines for Providing Appropriate Access to Geospatial Data in Response to Security Concerns." The guidelines include procedures to help identify sensitive information content in geospatial data sets and helps data producers provide appropriate access to the data while protecting sensitive content.

The Freedom of Information Act, 5 U.S.C. § 552, referred to as FOIA, stipulates that the United States Government, including the National Park Service, must provide access to data and information of interest to the public. FOIA, as amended in 1996 to provide guidance for electronic information distribution, applies to records that are owned or controlled by a federal agency, regardless of whether or not the federal government created the records. FOIA is intended to establish a right for any person to access

federal agency records that are not protected from disclosure by exemptions. Under the terms of FOIA, agencies must make non-protected records available for inspection and copying in public reading rooms and/or the Internet. Other records, however, are provided in response to written requests through a specified process. The Department of the Interior's revised FOIA regulations and the Department's Freedom of Information Act Handbook can be accessed at http://www.doi.gov/foia/ for further information.

In some cases, public access to data can be restricted. The National Park Service is directed to protect information about the nature and location of sensitive park resources under Executive Order No. 13007: Indian Sacred Sites, Director's Order #66: Freedom of Information Act and the Protection of Exempted Information (Draft, 12/10/2004), and four resource confidentiality laws:

- National Parks Omnibus Management Act (NPOMA; 16 U.S.C. 5937)
- National Historic Preservation Act (16 U.S.C. 470w-3)
- Federal Cave Resources Protection Act (16 U.S.C. 4304)
- Archaeological Resources Protection Act (16 U.S.C. 470hh)

Through these regulations, information that could result in harm to natural resources can be classified as 'protected' or 'sensitive' and withheld from public release (NPOMA).

According to NPOMA, if the NPS determines that disclosure would be harmful, information may be withheld concerning the nature and/or specific locations of the following resources recognized as 'sensitive' by the NPS:

- Endangered, threatened, rare, or commercially valuable National Park System resources (species and habitats)
- Mineral or paleontological objects
- Objects of cultural patrimony
- Significant caves

The following guidance for determining whether information should be protected is suggested in draft Reference Manual RM 66B: Handling Protected Information (Draft, 11/9/2004):

- Has harm, theft, or destruction occurred to a similar resource on federal, state, or private lands?
- Has harm, theft, or destruction occurred to other types of resources of similar commercial value, cultural importance, rarity, or threatened or endangered status on federal, state, or private lands?
- Is information about locations of the park resource in the park specific enough so that the park resource likely could be found at these locations at predictable times now or in the future?
- Would information about the nature of the park resource that is otherwise not of concern permit determining locations of the resource if the information were available in conjunction with other specific types or classes of information?
- Even when relatively out-dated, is there existing information that would reveal locations or characteristics of the park resource such that the information could be used to find the park resource as it exists now or is likely to exist in the future?
- Does NPS have the capacity to protect the park resource if the knowledge about its specific location is easily obtainable?

Note that information already in the public domain, in general, has no further restrictions on its release. For example, the media has reported in detail the return of condors to the Grand Canyon. If an individual requests site-specific information about where the condors have been seen, this information may be released. However, the locations of specific nest sites cannot be released.

Managing natural resource information that is sensitive or protected requires:

- Identification of potentially sensitive resources
- Compilation of all records relating to those resources
- Determination of what data must not be released to the public
- Management and archival of those records to avoid their unintentional release

Classification of sensitive I&M data will be the responsibility of the Sonoran Desert Network staff, the park superintendents, and project leaders working on individual projects. Network staff will classify sensitive data on a case-by-case, project-by-project basis. They will work closely with project leaders to ensure that potentially sensitive park resources are identified and that information about these resources is tracked throughout the project.

The network staff is responsible for identifying all potentially sensitive resources to individual project leaders. The project leaders, whether network staff or partners, will develop procedures to flag all potentially sensitive resources in any products that result from the project, including documents, maps, databases, and metadata. When submitting any products or results, project leaders should specifically identify all records and other references to potentially sensitive resources. Note that partners should not share or release any information before consulting with network staff to ensure that the information is not classified as sensitive or protected.

9.2.3 Access Restrictions on Sensitive Data

The Sonoran Desert Network staff is responsible for managing access to sensitive data handled by the Program. All potentially sensitive park resources will be identified and investigators working on network projects will be informed that:

- All data and associated information must be made available for review by network staff prior to release in any format.
- Any information classified as protected should not be released in any format except as approved in advance by the National Park Service.

The network coordinator, NPS project liaison, or data manager identifies all potentially sensitive park resources to the leader for each project. Reciprocally, the project leaders must identify any known references to potentially sensitive park resources that appear in any products resulting from the project. The network staff provides a complete list of all references to potentially sensitive park resources to each park superintendent for review. The superintendent then determines which information should be protected.

When preparing or uploading information into any network database, the network staff ensures that all protected information is properly identified and marked. The network staff works together to ensure that all references to protected information are removed or obscured in any reports, publications, maps, or other public forum. A certification checklist will be completed by the project leader and network staff prior to posting data and information.

Network staff will remove any sensitive information from public versions of documents or other media. They will isolate sensitive from non-sensitive data and determine the appropriate measures for withholding sensitive data. The main distribution applications and repositories developed by the I&M Program (see section 9.2.1) are maintained on both secure and public servers, and all records that are marked ‘sensitive’ during uploading will only become available on the secure servers. Procedures for

assigning a sensitivity level to specific records when uploading to both the NPSpecies and NatureBib databases are discussed at the following websites:

- http://science.nature.nps.gov/im/apps/npspp/index.htm
- http://www.nature.nps.gov/nrbib/index.htm

It is crucial that network staff institute quality control and quality assurance measures to ensure that the person doing the uploading of records into the online applications is familiar with the procedures for identifying and entering protected information. Thus, access to data on sensitive park resources can be limited to NPS/network staff or research partners. However, limits to how these data are subsequently released must also be clearly defined.

9.2.4 Public Access to Network Inventory and Monitoring Data

According to FOIA (specifically the 1996 amendments), all information routinely requested must be made available to the public via reading rooms and/or the Internet. Network project data will be available to the public at one or more of the following internet locations:

- Sonoran Desert Network website (http://science.nature.nps.gov/im/units/sodn/)
- Public servers for the NPSpecies and NatureBib databases
- Public server for the Biodiversity Data Store
- Public server for the NR-GIS Data Store

The Network will regularly provide updated information about inventories and monitoring projects, including annual reports and detailed project reports, through the network website. Information on species in the National Parks, including all records generated through the I&M Program, will be maintained and made accessible through the NPSpecies database. Bibliographic references that refer to National Park System natural resources will be accessible through the NatureBib database. Documents, maps, and data sets containing resource information from all sources, and their associated metadata, will be accessible through the Biodiversity Data Store and/or the NR-GIS Data Store. Each of these databases/repositories will be available via both a secure server and a public server, and the public can access all information in these databases except those records marked as 'sensitive.'

9.2.5 Data Availability

Both raw and manipulated data resulting from the network's inventory and monitoring projects will be fully documented with FGDC-compliant metadata and made available to the public. The metadata for all data sets (excluding protected information) will be made accessible to the public as soon as they are verified and certified by the project leaders and the data manager.

Data sets for short-term studies (*e.g.*, inventories) will be provided to the public on the SODN website two years following the year the data were collected or following publication of the researcher's results (whichever comes first). Long-term (monitoring) study results will be provided to the public in five-year intervals, or when trend analyses have been completed and reported on by the Network. (This will be specific to each network monitoring protocol; refer to the Network's Vital Signs Monitoring Plan for further information). Before data are posted, the project leader will be asked to verify the final data set and metadata. Once network staff and the project leader verify the data set, the data will be made accessible to the public, provided no sensitive information is identified.

Network staff will notify investigators prior to making data sets publicly available to allow each researcher the opportunity to request in writing further restrictions on public access to the data set.

Network staff will review the request and determine whether the request will be granted and for how long the data set will remain restricted.

9.2.6 Data Acquisition Policy

The Network will develop a data set acquisition policy that will be made available on the Network website for all users who wish to acquire program data and information not available through public interfaces. This policy will include:

- A mandatory questionnaire available on the website. This questionnaire must be completed and mailed to the network data manager before data can be acquired. (This questionnaire will allow network staff the ability to maintain a distribution log specifying recipient name and contact information, intended use of data, export file format, delivery date and method, and data content description noting range by date and geography of data delivered.)
- A statement about use and appropriate citation of data in resulting publications.
- Request that acknowledgement be given to the National Park Service Inventory and Monitoring Program within all resulting reports and publications.

Metadata for data sets that can be acquired from publicly accessible sites will include the last two items from the aforementioned Network acquisition policy.

9.3 Data Feedback Mechanisms

The network website will provide an opportunity for NPS staff, cooperators, and the public to provide feedback on data and information gathered as part of the SODN I&M Program. A 'comments and questions' link will be provided on the main page of the site for general questions and comments about the network's program and projects. A more specific 'data error feedback' link will direct comments pertaining to errors found in website-accessible data to the data manager. Annual reporting of progress will be presented to the Board of Directors and to the Technical Committee, and feedback will be expected during and following these presentations.

9.3.1 Data Error Feedback Procedures

The following feedback procedures describe the process that we will use to receive, evaluate, and verify data errors reported by public and private data users:

- Web users send in a notification about an alleged error through the network website. Network staff then sends an acknowledgment to the notifier.
- Network staff inputs the information into a data error log table incorporated in either each of the network monitoring databases or a specific error tracking database developed for the Network.
- Network staff determines if the data questioned by the notifier are correct or incorrect. If the data are correct, then the network staff informs the notifier that no corrections are to be made and the information stands. If the data are incorrect, the network staff makes the appropriate corrections and notifies the original data collectors (cooperator, other agency, park staff, etc.).
- When data are corrected, the network website will be refreshed with the corrected information.
- Throughout this process, the network staff will inform the notifier via email of the status.

Credits

This chapter was adapted from material developed by Sara Stevens (Northeast Coastal and Barrier Network).

Chapter 10. Data Maintenance, Storage, and Archiving

Effective long-term data maintenance depends on thoughtful and appropriate data documentation. An essential part of any archive is its accompanying explanatory materials (Olson and McCord 1998). This chapter will refer to, and in some cases elaborate on, metadata standards and data set documentation procedures that are more fully explained in Chapter 7 (Data Documentation) of the SODN Data Management Plan.

Data, documents, and any other products that result from projects and activities that use network data are all crucial pieces of information. To ensure high-quality long-term management and maintenance of this information, the Network will implement procedures to protect information over time. These procedures will permit a broad range of users to easily obtain, share, and properly interpret both active and archived information.

10.1 Digital Data Maintenance

In general, digital data maintained for a long time will be one of two types:

- *Short-term data sets*, for which data collection and modification have been completed (*e.g.*, inventory projects)
- *Long-term monitoring data sets*, for which data acquisition and entry will continue indefinitely

Following the lead of the National Park Service and the National Inventory & Monitoring Program, the Network has adopted MS Access as its database standard and ArcGIS as its spatial data management standard. The Network will remain current and compatible with National Park Service and National Inventory and Monitoring Program version standards for these software programs.

Technological obsolescence is a significant cause of information loss, and data can quickly become inaccessible to users if they are stored in out-of-date software programs, on outmoded media, or on deteriorating (aging) media. Effective maintenance of digital files depends on the proper management of a continuously changing infrastructure of hardware, software, file formats, and storage media. Major changes in hardware can be expected every 1-2 years and in software every 1-5 years (Vogt-O'Connor 2000). As software and hardware evolve, data sets must be consistently copied onto fresh media, migrated to new platforms, or saved in formats that are independent of specific platforms or software (*e.g.*, ASCII delimited text files).

Any data set for which data entry or updates are still occurring will be stored in an active projects directory on the Network's active file server. This includes data sets for short-term projects as well as the current year's non-certified data for a long-term monitoring project. An archive projects directory will store data sets that will no longer change such as the master database containing certified data for a long-term monitoring project.

10.1.1 Short-Term Data Sets

Upon project finalization, the project leader and the data manager are responsible for creating a set of ASCII comma-delimited text files for all short-term data sets created or managed by the Network. Each database table will have a corresponding ASCII file. These files will be accompanied by a README.TXT file that explains the contents of each file, file relationships, and field definitions. These ASCII files are in addition to the native version of the data set (typically in database format). Creating

these text files will help ensure that the data are usable in a wide range of applications or platforms. All finalized files will be stored on the network archive server in the appropriate project folder.

In addition to creating ASCII files, we will also update completed and archived data sets that may have been created in older versions of MS Access, with the goal of having no data set more than two versions behind the current version used by the Network. There is a risk of reducing performance in the process of conversion; for example, complex data entry forms or reports may not function properly in an upgraded version. To the extent possible, we will maintain proper functionality of data entry forms and reports. However, the priority will be to ensure basic table and relationship integrity. All previous versions of the data set will be saved.

10.1.2 Long-Term Monitoring Data Sets

Long-term monitoring data sets require regular updates and conversion to current database formats. All active or long-term databases will conform to the current NPS and I&M software version standards.

Monitoring projects will also have variable long-term data archiving requirements. Raw data sets that are later manipulated or synthesized may need to be stored in perpetuity. Modifications to protocols will typically require complete data sets to be archived before modifications are implemented. Depending on the monitoring project, it may be necessary to preserve interim data sets (data 'milestones') over the long term. We will save archived data sets or subsets destined for long-term archiving, whenever possible, in their native formats in addition to ASCII text files. Data archiving requirements for ongoing projects will be spelled out in the data management SOPs for each monitoring project.

10.1.3 Quality Control of Converted Data

All ASCII files created from databases will undergo quality control (QC) to ensure that the number of records and fields correspond to the source data set and that conversion has not created errors or data loss. A second reviewer (preferably a program scientist) will evaluate the ASCII files and documentation to verify that tables, fields, and relations are fully explained and presented in a way that is useful to secondary users.

Databases that are converted from one version of MS Access to an upgraded version will require additional QC. This is especially important if the databases are being actively used for data entry or analysis. Forms, queries, reports, and data entry all will be thoroughly tested.

10.1.4 Version Control

Previous versions of databases will be saved in their native format and archived in addition to the current version. Documentation of version updates and associated details will be part of the archive metadata document, and revision information and history will also be included in tables within the database files themselves. File names of the archived revisions should clearly indicate the revision number or date.

10.1.5 Spatial Data

Spatial data sets that are essential to the Network will be maintained in a format that remains fully accessible by the current version of ArcGIS. ArcGIS has maintained compatibility with previous data formats, and while only shapefiles retained all functionality in ArcGIS 8.x, both shapefiles and coverages retain their functionality in ArcGIS 9.x. At this time, there is no practical way to save GIS data in a software or platform-independent format.

Both uncorrected and corrected GPS data will be archived in their native format in addition to the corresponding GIS files that are created.

Remote sensing data, including satellite images, are stored in their native format and in processed formats for use in image processing and GIS applications.

10.1.6 Digital Still Images

The Network expects that most images acquired with handheld cameras will be still images in digital Joint Photographic Experts Group (JPEG) format. Network staff and cooperators are encouraged to use digital camera resolution and quality settings that produce images with adequate quality at a reasonable file size. Project crew members normally collect more images than necessary for documenting a project. The project leader should select and submit only those images necessary to complement the other project data. The data manager will review images for quality and file size and may process the images to meet specifications. Digital images are stored and named according to the instructions in each monitoring protocol.

10.2 Storage and Archiving Procedures for Digital Data

Digital data need to be stored in a repository that ensures both security and ready access to the data in perpetuity.

10.2.1 Directory Structure for Electronic Archives

The Network relies on two Dell PowerEdge servers running the Windows 2003 Server operating system with a level-5 RAID (redundant array of independent disks) array for data storage. Both servers are located in the Network office and are managed by network personnel and off-site IT specialists (see Chapter 3). The PowerEdge 4600 functions as the archive file server, and the PowerEdge 2800 functions as the active (working) file server. Network personnel have full access to the active server, but write access to the archive server is limited to data management staff.

Figure 10.1 illustrates the directory structure for both active and archived data files, including GIS data. This directory structure was devised, in part, to accommodate two different backup schedules, one for rapidly changing files, the other for relatively static files.

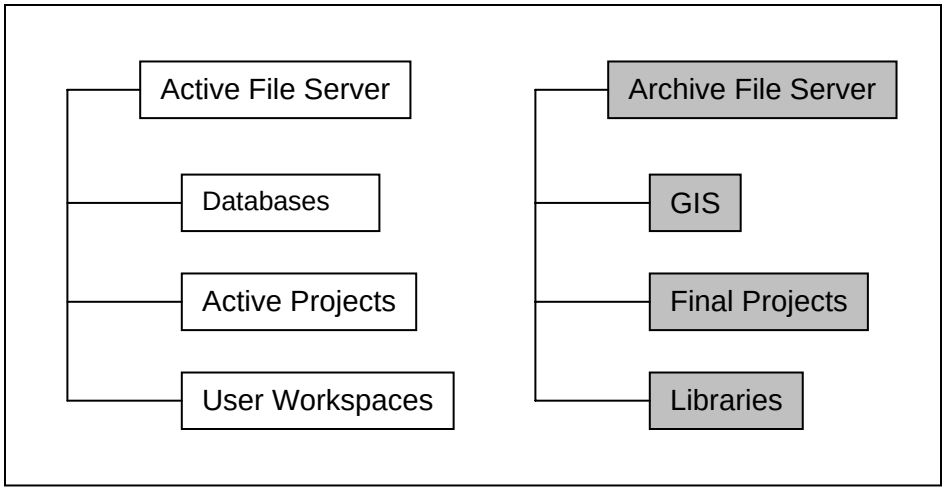

Figure 10.1. Current directory structure for data files. Note: All directories are not shown.

10.2.2 Directory Structure for Individual Projects

The organizing and naming of folders and files should be intuitive enough that users unfamiliar with a specific project can still easily navigate it. Since each project has its own variations and idiosyncrasies, a standardized structure is not realistic. However, all project archives will include several to most of the following elements:

- Administrative documents such as agreements, correspondence, and research permits
- Programmatic documents including protocols, procedures, and supporting documents
- Interim data sets or milestones
- Data sets submitted by contractors
- Data sets reformatted or manipulated by the Network (*e.g.*, data converted to NRDT format, data sets migrated to current software formats)
- Data sets in ASCII format
- Conceptual or statistical models used for data interpretation
- Final reports
- README files – including an explanation of directory contents, project metadata (including a Dataset Catalog report), and version documentation

Once final data and reports have been submitted, draft products do not need to be maintained.

10.2.3 Backup Procedures for Digital Data

The risk of data loss can come from a variety of sources, including catastrophic events (*e.g.*, fire, flood), user error, hardware failure, software failure or corruption, security breaches, and vandalism. Performing regular backups of data and arranging for off-site storage of backup sets are the most important safeguards against data loss.

Data stored on the active server is backed up onto five Lacie external hard drives, each designated for a weekly backup. A full backup of this server is run every Monday; incremental backups are performed every Tuesday-Friday of each week. Each hard drive backs up a full week and stores that data for four weeks. The archive server is backed up less frequently onto another external hard drive. Backups occur every two months or in response to significant changes to the data; *e.g.*, uploading a batch of new GIS data. This external hard drive is stored in a fireproof cabinet. All backups are performed and monitored by network personnel.

Table 10.1. Backup schedule for Network servers.

Schedule A – frequently-changing files. Used for all files located on SODN active server.	
Hard Drive 1	week 1: full backup and daily incrementals
Hard Drive 2	week 2: full backup and daily incrementals
Hard Drive 3	week 3: full backup and daily incrementals
Hard Drive 4	week 4: full backup and daily incrementals
Hard Drive 5	week 5: full backup and daily incrementals
Schedule B – large files, relatively static. Used for all files located on SODN archive server.	

Hard Drive 1	month 1: full backup month 3: full backup month 5: full backup

Backups of data stored on the personal computers of staff are the responsibility of each staff member. We strongly recommend that staff members store or regularly copy important files into their personal directory on the active server, where daily backups are performed. Staff may also place data in the archive section of the I&M server with the approval of the data manager.

This backup strategy will be re-evaluated at least annually to ensure it is meeting Program needs. Backup routines represent a significant investment in hardware, media, and staff time; however, they are just a small percentage of the overall investment that we make in program data.

10.2.4 Data and Network Security

Although the Network is not located within established National Park Services offices, local and wide area networks currently conform to Department of Interior security guidelines. Access to NPS IT resources is granted only to personnel who have completed the annual IT security awareness training.

Only I&M staff and system administrators have permission to access files on the network servers, and restrictions have been established on archived data files. Archive directories containing completed project data or certified data from ongoing projects are designated as read-only for all staff with the exception of the data manager. Therefore, any changes must be routed through the data manager, who is responsible for ensuring that documentation and README files associated with the data set are also updated.

10.3 Storage and Archiving Procedures for Documents and Objects

The guidelines in this section apply to documents such as final reports prepared by staff or contractors, program administrative documents, contracts and agreements, memoranda of agreement or understanding, and other documents related to network administration, activities, and projects. These guidelines also apply to physical items such as natural history specimens, photographs, or audio tapes. In most instances, these documents and objects are essential companions to the digital data archives described earlier.

Direction for managing these materials (as well as digital materials) is provided in NPS Director's Order 19: Records Management (2001) and its appendix, NPS Records Disposition Schedule (NPS-19 Appendix B, revised 5-2003). NPS-19 states that all records of natural and cultural resources and their management are considered mission-critical records. NPS-19 further states:

> "Mission critical records are permanent records that will eventually become archival records. They should receive the highest priority in records management activities and resources and should receive archival care as soon as practical in the life of the record."

Section N of Appendix B, which provides guidelines on natural resource-related records (including, specifically, the results of Inventory and Monitoring Programs), indicates that all natural resource records are considered 'permanent'; *i.e.*, they are to be transferred to the National Archives when 30 years old. It also indicates that non-archival copies of natural resource-related materials are "potentially important for the ongoing management of NPS resources" and should not, in any instance, be destroyed.

10.3.1 Documents

All paper documents managed or produced by the Network will be housed in one or more of three locations:

Network central files, Tucson, Arizona

These files contain project files, administrative documents, and non-record copies of documents that are archived at an off-site facility (see item 2, below). Examples include meeting minutes, correspondence, memoranda of understanding, contracts and agreements, research permits, and interim and selected final reports produced by the program or under its auspices. We will use acid-free paper and folders for all permanent records in the central files. In addition to maintaining these paper records, we will maintain electronic versions, when possible, on the server. The central files are maintained by network data management personnel under the guidance of the data manager and network coordinator.

Western Archeological and Conservation Center (WACC), Tucson, Arizona

[Repository subject to change]. WACC provides temperature and humidity-controlled facilities, a professional archival staff, and meets all museum standards set by the NPS. This repository will be used for original documents and associated materials produced by the Network (*e.g.*, photographs, field notes, permits) that are a high priority to maintain under archival conditions. Examples include original inventory reports and accompanying slides and maps; original vegetation mapping reports; and SODN Monitoring Plans. Copies of these reports will be maintained in the Network central files, and all will have an electronic equivalent (*e.g.*, pdf) for distribution or reproduction.

For all materials submitted to WACC, the Network will provide essential cataloging information such as the scope of content, project purpose, and range of years, to facilitate ANCS+ record creation and accession. We will also ensure that materials are presented using archival-quality materials (*e.g.*, acid-free paper and folders, polypropylene or polyethylene slide pages or photo/negative sleeves). Upon the recommendation of museum staff, the Network will use Light Impressions (www.lightimpressionsdirect.com) as the source for most of its archival storage materials.

Many Network reports and documents encompass data from multiple parks, which has made it difficult to accession archival copies into a specific network park museum. In these instances, WACC will prepare associated ANCS+ records that reference all parks included in a report or document, and they will create finding aids to help potential users locate the materials.

Network Parks' Central Files and/or Museums

The Network will provide high-quality copies of park-related documents resulting from I&M projects, along with electronic versions, to park resource management staff. Parks may choose to accession these materials into their museums, incorporate them into their central files, or house them in their resource management library, as they deem appropriate. The Network will not manage documents at the park level.

10.3.2 Specimens

The Network will provide specimens collected under the auspices of the SODN I&M Program to the network park in which they were collected for curation, or to a repository approved by a park (where the specimens are considered on loan). We will provide park curators with associated data required for cataloging each specimen. These data will be in comma-delimited format (.csv) format for automated uploading into ANCS+. Data will be provided to non-NPS curators in Excel format.

10.3.3 Photographs

Archivists have been reluctant to fully embrace digital photography, and some have expressed concern that, with the acceleration of technological change, documentary heritage is in danger of being lost in the information age (Cox 2000).

The Network has chosen to take a conservative approach and requires staff and contractors to provide photos as 35mm slides (preferably Kodachrome or Ektachrome), which have a proven long-term stability (Wilhelm and Brower 1993). If contractors cannot provide slides, the Network requests 4x6 color prints. [This policy is under review.] Original images are a high priority for placing in archival storage conditions.

Slides should be labeled using indelible pigment ink or laser-printed archival-quality slide labels. Slide labels should include a unique ID, project name, photographer, photo date, a brief identification of contents (*e.g.*, species name, plot ID), and geographic location (UTMs or description). All slides will be stored in polypropylene slide sleeves at the SODN office until transferred to WACC. In addition, all slides will be scanned and saved as TIFF files, and these electronic copies will be used as the primary means of distributing or reproducing the images.

If photographs are provided, they will be stored in individual polypropylene sleeves within archival boxes. Each photo will be labeled on the back, using archival-quality labels that are either laser-printed or handwritten, with the same information elements required for slides. If a partner is submitting photographs, corresponding TIFF files must also be submitted.

Every image, regardless of format, will be entered into the SODN Photo Database, where attributes such as electronic file name, keywords, project, photo description, photographer, date, and location will be cataloged. All digital photos are housed in the Libraries section of the archive data server (see Figure 10.1).

10.3.4 Role of Curators in Storage and Archiving Procedures

Curators for parks within the Network are an ongoing source of expertise, advice, and guidance on archiving and curatorial issues, and they have a role in almost every project undertaken by the Network. Project leaders should involve park curators when projects are in the planning stage to ensure that all aspects of specimen preservation or document archiving will be considered and that any associated expenses are included in project budgets.

Credits

This chapter was adapted from material developed by Margaret Beer (Northern Colorado Plateau Network).

Chapter 11. References

Brunt, J. W. 2000. Data management principles, implementation and administration. Pages 25-47 *in* W. K. Michener and J. W. Brunt, editors. Ecological data: Design, management and processing. Blackwell Science Inc., Malden, MA.

Chapal, S. E., and D. Edwards. 1994. Automated smoothing techniques for visualization and quality control of long-term environmental data. Pages 141-158 *in* W. K. Michener, and J. W. Brunt, and Susan G. Stafford, editors. Environmental information management and analysis: Ecosystem to global scales. Taylor & Francis Ltd., London.

Cox, R. J. 2000. Searching for authority: archivists and electronic records in the New World at the fin-de-siècle. First Monday **5**:1. http://firstmonday.org/issues/issue5_1/cox/index.html.

Davis, K., and W. L. Halvorson. 2000. A study plan to inventory vascular plants and vertebrates: Sonoran Desert Network, National Park Service.

Edwards, D. 2000. Data quality assurance. Pages 70-91 *in* W. K. Michener and J. W. Brunt, editors. Ecological data: Design, management and processing. Blackwell Science Inc., Malden, MA.

Gregson, J. 2000 Draft. Natural resources data and information systems handbook outline.

Gregson, J. 2002 Draft. Natural resource data and information systems management handbook: Chapter 2. http://www1.nrintra.nps.gov/im/datamgmt/nrdismhb.htm. Accessed 15 September 2004.

Mau-Crimmins, T., A. Hubbard, D. Angell, C. Filippone, and N. Kline. September 2005. Sonoran Desert Network Vital Signs Monitoring Plan. Technical Report NPS/IMR/SODN-003. National Park Service. Denver, CO.

Michener, W. K. 2000. Metadata. Pages 92-116 *in* W. K. Michener and J. W. Brunt, editors. Ecological data: Design, management and processing. Blackwell Science Inc., Malden, MA.

Michener, W.K., and J.W. Brunt, editors. 2000. Ecological data: Design, management and processing. Blackwell Science Inc., Malden, MA.

Miller, A. B. 2001. Managing data to bridge boundaries. Crossing Boundaries in Park Management: Proceedings of the 11th Conference on Research and Resource Management in Parks and on Public Lands, The George Wright Society, Hancock, MI, 2001:316-320.

National Park Service. 2002. Director's Order #11B: Ensuring quality of information disseminated by the National Park Service. http://www.nps.gov/policy/DOrders/11B-final.htm. Accessed 15 September 2004.

National Park Service. Draft 2004. Director's Order #66: Freedom of Information Act and the Protection of Exempted Information. http://www1.nrintra.nps.gov/DO66/DO_66_Final_Draft.doc. Accessed 14 December 2004.

National Park Service. 2003. NPS records disposition schedule. NPS-19, Appendix B – Revised, 5-03.

National Park Service. Draft 2004. Reference Manual RM-66A: FOIA processing. http://www1.nrintra.nps.gov/DO66/RM-66A-Draft.doc. Accessed 14 December 2004.

National Park Service. Draft 2004. Reference Manual RM-66B: Handling protected information. http://www1.nrintra.nps.gov/DO66/RM-66B-Draft.doc. Accessed 14 December 2004.

Olson, R. J., and R. A. McCord. 1998. Data Archival. Pages 53-57 *in* W. K. Michener, J. H. Porter, and S. G. Stafford, Data and information management in the ecological sciences: A resource guide. LTER Network Office, University of New Mexico, Albuquerque.

Palmer, C. J. 2003. Approaches to quality assurance and information management for regional ecological monitoring programs. Pages 211-225 *in* D. E. Busch and J. C. Trexler, editors. Monitoring ecosystems: interdisciplinary approaches for evaluating ecoregional initiatives. Island Press, Washington, DC.

Palmer, C. J., and E. B. Landis. 2002 Draft. Lake Mead National Recreational Area Resource Management Division: Quality system management plan for environmental data collection projects. http://www1.nrintra.nps.gov/im/datamgmt/docs/RMQSMP_2.pdf. Accessed 15 September 2004.

Tessler, S., and J. Gregson. 1997. Draft Data Management Protocol. http://www1.nrintra.nps.gov/im/dmproto/joe40001.htm. Accessed 15 September 2004.

U.S. Environmental Protection Agency. 2003. Guidance for geospatial data quality assurance project plans. Office of Environmental Information, Washington, DC. EPA QA/G-5G.

Vogt-O'Connor, D. 1997. Caring for photographs: general guidelines. National Park Service Conserve O Gram **14**:4. NPS Museum Management Program, Washington, DC.

Vogt-O'Connor, D. 2000. Planning digital projects for preservation and access. National Park Service Conserve O Gram **14**:4. NPS Museum Management Program, Washington, DC.

Wilhelm, H., and C. Brower. 1993. The permanence and care of color photographs: Traditional and digital color prints, color negatives, slides, and motion pictures. Preservation Publishing Company, Grinnel, IA.